Synthesizing Conversational Intonation from a Linguistically Rich Input

Paul Taylor[1] and Alan W. Black[2]

(1) Human Communication Research Centre, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW. email: paul@cogsci.ed.ac.uk

(2) ATR Interpreting Telecommunication Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, JAPAN. email: awb@itl.atr.co.jp

Abstract

This paper describes a general system which maps from a phonological specification of an utterance's intonation to a F_0 contour. The system can accommodate a variety of feature based phonological description schemes. Speaker dependent characteristics can be modelled and an automatic method of determining these is described.

1 Introduction

This paper describes the phonetic component of the intonation module of the CHATR speech synthesizer developed at ATR. CHATR is a concept-to-speech system which is used as the final stage in a spoken language translation system [2]. The input to CHATR is not text-based, but rather a complex linguistic description of an utterance that is produced from the language generation part of the translation system, and so reliable syntactic, semantic and pragmatic information is available for each utterance to be synthesized. The utterances are part of a conversation, so it is not acceptable to produce "neutral declarative" intonation: a much wider range of intonational effects must be modelled. It is also an important long-term goal to be able not only to translate and synthesize the spoken words of a user, but also to model the speaker dependent characteristics of a user's voice.

The linguistic component of the intonation module takes input from the language generation system and determines prosodic phrasing, pitch range, pitch accent locations and pitch accent type. The details of this component are described in Black and Taylor [1]. The phonetic component takes this information and produces F_0 contours. Figure 1 shows the layout of the system.

2 Intonational Events

The input to the phonetic component is a list of prosodic phrases, each containing a start F_0 value and a list of syllables. The start F_0 value represents pitch range and is given as a single normalised number representing where in the speaker's pitch range the first F_0 value of the phrase lies. Syllables contain a list of segment descriptions, each of which has a name for the segment and a duration (previously derived from the duration module [3]). Syllables can optionally have one or more *intonational events* associated with them.

An intonational event is a general term for a phonologically significant intonational effect. Pitch accents, declination resets and phrase-final boundary rises are all intonational events. The processing of the phonetic part of the prosodic component is mainly concerned with converting the linguistically based event descriptions of the input into acoustically relevant descriptions suitable for waveform synthesis. Between every event, there is a *connection*. Connections represent the parts of contours where there is nothing of intonational significance.

An event based model is helpful as it gives us a clean generic mechanism which we can use to model all important intonational phenomena. In the phonetic component, no *computational* distinction is made between pitch accents, declination resets and boundary rises, and a single processing paradigm is used throughout.

3 From Features to Tilt

3.1 Feature Definitions

Initially events and connections are described using features. A *feature definition table* describes which features can be used, and this is configurable. Table 1 shows the particular feature set that is used at present.

Each feature in the table has a name which serves as an identifier, a field indicating if the feature operates on events or connections, and a type, which is either *binary* or *scalar*. Binary features are either present or not present, whereas scalar are always present and take a continuous value. Thus each event in the input is marked with the binary features which are present, and a value for each scalar feature.

The first module in the phonetic component takes each event from the input utterance, and converts the features descriptions into descriptions in the *tilt* space. Each event in the tilt space is described by four continuous parameters. *Amplitude* represents the size of the event in frequency terms, *duration* is the length of the event and *position* shows the alignment between the event and

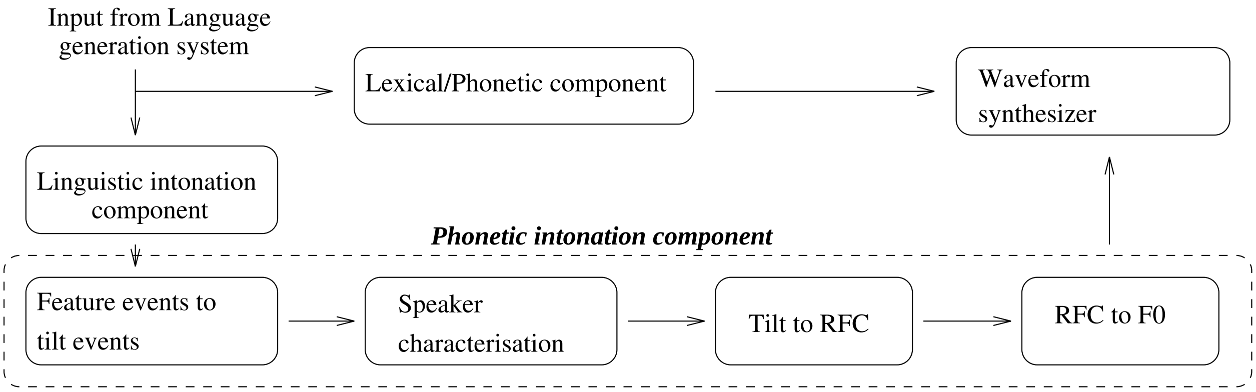

Figure 1: *System diagram showing processing from feature based events to F_0 contours*

the segments. *Tilt* is a continuous quantity which describes the shape of the event. Tilt values are dimensionless and range between 1 and -1. A tilt value of 1 is an event where the F_0 contour only rises, a value of -1 means that for the duration of event the F_0 contour is falling, and a tilt value of 0 indicates the event has a rise and a fall of equal size (the precise definition is given in the next section). In the tilt space, connections are represented by a single parameter, *amplitude*.

The feature definition table also specifies how each tilt parameter is affected by each feature. Often a feature has no effect on a particular parameter and in that case the entry in the table is left blank. If a definition is given, it is in the form of an *operator* and a *value*. The operator can be one of three types: *assignment (=)* means that the event parameter is given the value in the table; *plus-assignment (+=)* means that value in the table is added to the value the event parameter already has; *multiply-assignment (*=)* means that value in the table is multiplied by the value the event parameter already has. It is important to note that if all the feature definitions use the same operator, the feature system is unordered (i.e. commutative). If different operators are used, the same features applied in a different order may have a different effect.

3.2 Speaker Characterisation

The values in the feature definition table represent the z-scores of normalised quantities, thus an amplitude of 0.0 indicates an average amplitude, and an amplitude of 1.0 represents an amplitude of 1.0 standard deviations greater than the average. To convert these values into real quantities, a *speaker characterisation table* of means and standard deviations is used. A different set of means and standard deviations is used for each speaker. These values are collected using the method described in section 5.1. The speaker characterisation table also contains the mean and standard deviation of the phrase initial F_0 value for each speaker.

For each event, the z-score value of each parameter is multiplied by the standard deviation and then added to the mean. After this, each event's amplitude is specified in terms of Hertz and the duration and position are specified in terms of milli-seconds. The tilt parameter is not normalised and is therefore not affected by characterisation. The connection amplitudes and start F_0 values for each phrase are also converted into Hertz values.

4 From Tilt to RFC space

The next operation is to convert each event's description into a form more amenable to F_0 calculation. The events and connections in tilt space are now converted into *rise/fall/connection* (RFC) space. The RFC model is fully described in Taylor [4], but a brief description is useful here. In the RFC space, each event is described by a *rise element* followed by a *fall element*. Each element has an amplitude and duration. An event with a tilt of -1 has a zero rise amplitude and an event with a tilt of +1 has a zero fall amplitude giving in effect rise-only and fall-only events.

The relation between the tilt, duration and amplitude parameters in the tilt space and the rise and fall amplitude parameters in the RFC space is shown below (the position parameter is unchanged).

$$D_{rise} = \frac{D(1+tilt)}{2} \quad D_{fall} = \frac{D(1-tilt)}{2} \tag{1}$$

$$A_{rise} = \frac{A(1+tilt)}{2} \quad A_{fall} = -\frac{A(1-tilt)}{2} \tag{2}$$

Once the events have been transformed into RFC space it is possible to generate an F_0 contour. The durations of the connection elements are calculated using the position parameter, the durations of the rises and falls and the durations of the segments. The start F_0 value for each phrase is given in the input, and the rest of the contour is created by processing the list of rise, falls and connection left-to-right and applying equation 3 for rises and falls and a straight line equation for the connections.

$$\begin{array}{ll} f_0 = A - 2.A.(t/D)^2 & 0 < t < D/2 \\ f_0 = 2.A.(1 - t/D)^2 & D/2 < t < D \end{array} \tag{3}$$

where A is element amplitude, D is element duration and t is time.

5 Configuring the System

Before the system can be used in practice, the speaker characterisation and feature definition tables must be defined.

5.1 Automatic Analysis

Algorithms have also been designed which can convert F_0 contours into a list of feature based events. This analysis facility is useful as it allows the collection of speaker specific data for the speaker characterisation procedure. In addition, it allows us to produce utterances labelled with feature based events, which helps in the design of the feature definition table and of the rules in the linguistic prosodic component which must generate the feature based events descriptions.

The most difficult part of the analysis process is getting an RFC based event description from the F_0 contour. This procedure is described in Taylor [4], but is as yet only about 75% reliable on speaker independent open test data and thus some hand editing is required.

The transformation from an RFC event to a Tilt event is straightforward. Tilt duration is the sum of the rise and fall durations, tilt amplitude is the sum of the *absolute* rise and fall amplitudes, and position is unchanged. Initially, there are two tilt parameters, one for amplitude and one for duration. The equations for these are given in 4 and 5.

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{4}$$

$$tilt_{dur} = \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}} \tag{5}$$

Experimental evidence has shown that amplitude and duration tilt are very strongly correlated, and this enables us to combine both these variables in a single tilt variable, which is the average of the two. Events in tilt space can be described using one less parameter than events in RFC space because the tilt parameters are in general more orthogonal. Because of this greater orthogonality in the tilt space, it is easier to design feature sets, as feature operations are often naturally orthogonal themselves.

Normalisation is the analysis equivalent of the characterisation procedure outlined in section 3.2. Once a significant number of events from a speaker have been labelled in the tilt space, it is possible to calculate the mean and standard deviation for each of the four parameters. The parameters for each event are normalised by subtracting the mean and dividing by the standard deviation.

Features are labelled using an analysis by synthesis approach. An artificial event is synthesized for every possible binary feature combination. The feature combination whose artificial event most closely matches the original is chosen. Scalar features are simply assigned the same value as in the original event. The feature labelling algorithm effectively quantizes the events in the tilt space.

5.2 Feature Sets

There are an infinite number of possible feature sets that can be tried in the system, so choosing a particular set is not trivial. It should be clear that as one uses less features, the tilt space will be quantized more heavily, and hence the synthesized contours will have poorer resolution. Conversely, the more features that are used, the smaller the quantization error will be. Continuous features in particular are very powerful, and a maximum of four are needed to synthesize any event in the tilt space (one feature for each tilt parameter). It should be remembered that the higher level parts of the prosodic component have to generate these feature based descriptions, and it is in the interest of those modules to have as simple a description system as possible. Thus there is a trade off between wanting a rich feature set which gives good synthesis quality and a compact set which is easy to generate.

The design of feature sets is still the subject of ongoing research but the set currently used is shown in table 1. In this particular set up, each feature affects only one tilt parameter, though this

Feature name	type	operates on	affecting variable	operator	with value
rise	binary	event	tilt	+=	+1
fall	binary	event	tilt	+=	-1
late	binary	event	position	+=	+1
early	binary	event	position	+=	-1
amp	scalar	event	amp and duration	=	scalar
rise	binary	connection	amp	+=	+1
fall	binary	connection	amp	+=	-1

Table 1: *A feature definition table*

Speaker	feature / tilt	tilt/ RFC	RFC / original	feature / original
FALZ	13.4	5.25	4.1	16.0
FJMT	15.5	4.8	4.9	18.3
MAEM	8.5	4.8	3.6	11.1
MMAG	8.3	7.1	3.7	13.1

Table 2: *Accuracy of synthesis for four American speakers. The figures show the difference in Hertz between contours synthesized from the event type on the left of the "/" and those synthesized from the right side.*

need not be the case. In this table, +early, +late has the same effect as -early, -late, and likewise with rise and fall, thus there are only 9 types of event.

The early and fall features are often used to indicate low or downstepped pitch accents, and the late feature is used to indicate "scooped" or "rise-fall" pitch accents. The rise feature is nearly always used for boundary and declination reset events.

6 Data and Results

If an event in one space is mapped to an event in a different space and back again, a slight error will occur as none of the mappings are exact. One can assess the overall performance of the system by labelling F_0 contours in terms of the feature based event descriptions, synthesizing contours from these descriptions and comparing the synthesized and original contours. This gives a measure of how accurately the feature based events encode the F_0 contours. (But does not directly give a measure of how useful a particular feature definition table is for any other purpose). It is also possible to perform the mappings between the RFC space and the F_0 contour and the tilt space and the F_0 contour. In this way we can see which part of the system gives the greatest error.

The system (using the features in table 1) was tested on 96 utterances from two American male and two American female speakers from the CMU-ATR conference registration database. The results are shown in table 2. The largest source of synthesis inaccuracy comes from the feature to tilt part of the mapping.

Informal perceptual testing has shown that the differences between contours synthesized from RFC and the tilt events are usually imperceptible. Often a slight difference can be heard when comparing feature synthesized contour and originals.

7 Conclusion

The system presented here provides a computationally explicit framework for synthesizing F_0 contours from phonological specifications. The modular nature of the system allows the design of speaker universal feature sets without having to specify speaker dependent details. Automatic analysis makes it is possible to model the characteristics of a speaker's voice in a well defined manner. The feature labelling capability provides automatic classification of events, allowing easy testing of feature schemes.

References

[1] Alan W. Black and Paul A. Taylor. Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input. In *ICSLP '94, Yokohama, Japan*, 1994.

[2] Alan W. Black and Paul A. Taylor. CHATR: A generic speech synthesis system. In *COLING '94, Kyoto, Japan*, 1994.

[3] W. N. Campbell and S. D. Isard. Segmental durations in a syllable frame. *Journal of Phonetics*, 19:37–47, 1991.

[4] Paul A. Taylor. The rise/fall/connection model of intonation. *Speech Communication*, 1994. (forthcoming).