# Three Methods of Intonation Modeling

*Ann K. Syrdal*[1] *Gregor Möhler*[2†] *Kurt Dusterhoff*[3†] *Alistair Conkie*[1] *Alan W. Black*[3]

[1]AT&T Labs - Research, Florham Park, NJ, USA
[2]Institute of Natural Language Processing, University of Stuttgart, Stuttgart, Germany
[3]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, Scotland

## ABSTRACT

This paper compares different methods of generating intonation for an American English Text-to-Speech synthesis system. We look at a primarily rule-based approach and two data-driven approaches.

For data-driven modeling we used two separate data sets, each representing a somewhat different prosodic style. One database was recordings of a portion of 1989 Wall Street Journal text from the Penn Treebank Project. The second database was recordings of interactive prompts used in telephone network services. Both were read by the same female speaker. Approximately two and one-half hours of speech was phonetically and prosodically segmented and labeled (first automatically, and subsequently verified manually). The prosodic labeling used ToBI [7] tones and breaks. Three different intonation models were compared: (1) a predominantly rule-based model based on ToBI labels [3]; (2) a parametric model using the Tilt approach [8]; and (3) a Vector Quantized model based on an underlying parametric representation [5]. Sentences representative of both prosodic styles were synthesized with each of these models, and were presented to listeners for subjective ratings in a formal listening test. The results of the evaluation are reported.

## 1. INTRODUCTION

Intonation generation is a critical part of any Text to Speech (TTS) Synthesis system. Intonation is important for intelligibility, and critically important for naturalness. It has also been, historically, one of the less well developed components of most TTS systems. Functionally, intonation generation is often divided into two parts. The first is locating and characterizing accents. This may be considered as primarily a linguistic problem, since it depends in large part on the syntax and semantics of the sentence to be synthesized. The second functional part is assigning a suitable F0 realization, given these accent specifications. We have adopted this two-stage approach in the current study. In addition, for the purposes of this experiment, we do not investigate how the accent locations and accent types are determined, but assume them as givens. Our focus is on looking at various ways in which the purely symbolic accent representation can be used to generate an F0 contour, and on perceptually evaluating them.

There are of course a variety of ways in which prosody generation can be performed. We have chosen here to look at a selection of methods. Our goals are to be able to assess both the quality of the realization in terms of perceived synthesis quality, and also to quantify just how important the input information is to overall synthesis quality. The selection of methods is of course constrained by those we have had the opportunity to study and by our desire to compare the methods in similar contexts. Thus we have restricted ourselves to three representative and distinctly different systems, all of which are to some degree data-driven. They are described in detail in the subsequent sections. Here we simply summarize their salient characteristics.

- The first model is primarily a rule-based model [3] based on speech examples from the ToBI training materials. It uses a target interpolation scheme with accent and boundary markers that are ToBI labels. Targets are placed with reference to syllable structure, within a pitch range specified by top and base lines. These top and base lines are derived from our speaker's speech data using a CART tree.
- The second method uses the Tilt model. Tilt accents are estimated from the accents marked in the speech database and parameterized using a parabolic approximation. CART trees are trained to relate segmental and prosodic features to each of the Tilt parameters. For synthesis, the most appropriate set of parameters is predicted using these CART trees.
- The third method is a parametric intonation event model (PaIntE) [5]. This is also a data-based model. The parameter set and the approximation functions differ from Tilt. Furthermore the parametric intonation events are vector quantized. For synthesis a particu-

[†] Part of the work for this paper was carried out at AT&T Labs during visits by G. Möhler and K. Dusterhoff

lar vector is predicted and converted into an intonation event.

## 2. THE RULE-BASED SYSTEM

The rule-based approach used in this paper is based on an in-depth analysis of a ToBI labeled reference corpus [3]. It follows the basic principles of the tone sequence model by applying a target interpolation scheme, i.e. each accent or boundary represented by a ToBI label is realized by a number of discrete F0 targets that are connected by linear interpolation. The targets reflect the H and L atoms in the ToBI transcription and are therefore realized high or low in the speaker's pitch range. The pitch range is seen as the distance between baseline and topline, which encompass the F0 contour. Thus, the position of a target value is described in terms of fractions of the distance between the two extremes. The pitch range of an utterance is influenced by the global structure of discourse. Since we do not have access to such information, we derive the pitch range from simple features like pausal duration, phrase length and number of accents using CART technology. The temporal position of a target is also given in percentages in relation to the voiced part of the syllable. The complete rule-set contains about 60 rules that realize each ToBI accent and boundary in different phrasal contexts.

## 3. TILT

A Tilt labeling for an utterance consists of an assignment of one of four basic intonational events: pitch accents, boundary tones, connections, and silence (labeled a, b, c, sil). Each of the events includes a number of continuous parameters. All events have a *start* parameter for the fundamental frequency at the start of the event (measured in Hertz). Pitch accents and boundary tones are also described by a *duration* (seconds), an absolute *amplitude* (Hertz), the *peak position* at which the rising portion of the event stops and the fall begins (measured in seconds from the start of the vowel), and a *tilt* value representing the "tilt" of the accent (described below).

The tilt parameter represents the amount of fall and rise in the accent. The starting F0 of an event acts as a point from which all other calculations may be made. The absolute amplitude from the starting F0 to the peak is the first portion of the absolute amplitude parameter. The other portion is the absolute amplitude from the peak to the end of the event. Either of these portions may be zero, if the event is a simple rise or simple fall. The two amplitude values are added together to form the absolute amplitude value. The tilt parameter is the difference of the amplitudes divided by their sum [8].

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|}$$

The tilt parameter has a range of -1 to 1, where -1 is pure fall, 1 is pure rise, and 0 contains equal portions of rise and fall.

A tilt labeling can be automatically derived from an F0 contour. However the labeling better represents the contour when accent and boundary position are hand-identified ("a" and "b"). In this case the hand labeled ToBI accents were used to derive "a"s and "b"s while the parameters themselves were derived automatically. This modeling building technique is described in [2]

## 4. THE PARAMETRIC INTONATION EVENT SYSTEM

The Parametric Intonation Event (PaIntE) system is in many ways similar to the Tilt model. Both are data-based and attempt to model the accents by using a small set of continuously-varying parameters. The PaIntE model uses a representation which is a sum of two sigmoids a fixed distance apart in time, with five parameters as follows:

- a1 and a2 signify the steepness of the rising and falling sigmoids respectively.
- b represents the alignment of the function a value of 1 means the syllable's length has been chosen.
- c1 and c2 model the amplitudes of the rising and falling sigmoids.
- d is the F0 value of the function's peak.

### 4.1. THE VECTOR-QUANTIZED PaIntE SYSTEM

The six PaIntE parameters previously have been used directly for modeling and synthesis [4]. However an extra step was taken here, and Vector Quantization (VQ) was applied to the PaIntE parameters. This is motivated by several things. First intonation theory suggests that intonation can be described by a number of distinct *shapes*. Also, the task of modeling the data is made easier. Using VQ effectively means that the parameters can be calculated together rather than independently – the output is constrained to give sets of parameter values that can co-occur naturally.

When predicting accents the system can be configured to either use ToBI labels or accent placement data alone. The experiment used an example of both cases.

One parameter that is unspecified in the description above is the VQ codebook size. Values of 8 and 16 were judged to offer the best compromise between lower accuracy due to predicting elements of a large codebook and poorer synthesis due to not being able to model the details of accent F0 curves adequately. See [5] for some further details.

# 5. THE FESTIVAL SYSTEM

We describe the Festival Speech Synthesis System [1] insofar as it is relevant to the experiments described here. Festival is a complete TTS system in that it attempts to take arbitrary text input and synthesize it. It is also a useful research tool given the widespread availability, ease of configuration and the ability to add new external modules. The basic Festival system has the following major modules available: text preprocessing, lexical lookup, prosodic assignment and waveform synthesis. Each level is fully configurable and designed to allow new modules to be easily added. We have modified the intonation assignment modules and use our own synthesis component. In addition modules for all three intonation methods under discussion were installed in the system. In order to generate the stimuli for the listening tests the system was configured to use each of the three different F0 generation models, with the remainder of the synthesis process unchanged.

# 6. ACOUSTIC INVENTORY

One of the unusual aspects of the test is that part of the same database that was used for training prosody models was also used in the acoustic inventory for the Unit Selection-based synthesizer used to synthesize the test utterances.

The synthesis method requires that units are selected at run time from an acoustic inventory that consists of many units of a given type that may differ in some aspects (e.g. F0, duration, or context). The most appropriate units for synthesis are selected from the multiplicity of available units by determining (1) which most closely resembles the synthesis specification, represented numerically by a "target cost" value (a lower value representing a better match), and (2) how well selected units might fit together, represented by a "concatenation cost" value (again, a lower value representing a better match). The process of selecting from the many possible sequences of units, each with different sets of costs, is done by using a Viterbi search to determine the lowest cost sequence. This sequence is then synthesized.

The relevant point to note is that since there is an overlap between the acoustic inventory and the prosody database, there is a possibility that interesting interactions will arise. Specifically, if a prosody modeling technique is able to accurately model the prosody of an in-inventory utterance, there is a possibility that the units in the acoustic inventory that belong to that recorded utterance will be chosen, since they are clearly suitable units to make up the utterance. Conversely, if a prosody modeling technique does not model accurately the prosody of a particular sentence, there is a smaller possibility that the units will be chosen from the units that make up that utterance in the inventory.

If spoken sequences of units are chosen, the synthesis quality in general will be higher, which should be reflected in higher ratings by the listeners. Thus good prosody that accurately models the natural prosody of an in-inventory utterance should be rated highly for two reasons: (1) the prosody itself is natural-sounding, and (2) the units selected are likely to be optimal for the utterance. The interaction of prosody and unit selection may increase the sensitivity of our test somewhat by further differentiating the prosody models, and it emphasizes the importance of good prosody modeling to further maximize overall synthesis quality through the selection of the most appropriate units.

This issue is demonstrated and discussed further in the results section.

# 7. PERCEPTUAL EVALUATION

## 7.1. Prosodic Styles and Corpora

For all three prosody models, two prosodic styles were trained separately: a news-reading style was trained with the WSJ corpus, and an interactive prompt style was trained with the Prompt corpus. Distributions of accent types varied somewhat between the two prosodic styles. For example, there was a relatively higher proportion of H* pitch accents in the Prompts than in the WSJ corpora, and a relatively higher proportion of L+H* accents in the WSJ corpus than in the Prompts. The Prompts are typically described as having more lively F0 variation than the WSJ corpus, but this may not be simply quantified. Mean F0 was 10 Hz higher for the Prompts test utterances than for the WSJ test sentences. The prosody evaluation not only compared prosody models, but compared ratings of the two types of prosodic styles as well.

For each of the two corpora, 80% of the database was used only for training, and 20% of the database was reserved for testing only. The WSJ corpus was approximately 104 minutes long, and the Prompt corpus was approximately 41 minutes of speech.

## 7.2. Test Materials

Test utterances were sampled from the 20% test subset of the corpora. Twelve test utterances were chosen from the test subsets of the two corpora. Six utterances were from the Wall Street Journal (WSJ) corpus, and six from the Prompts corpus. They were chosen at random from utterances which were considered short enough for the purposes of a perceptual test (less than 10 seconds). All six WSJ test utterances were single sentences, although some were fairly complex. The mean number of words in WSJ test sentences was 17.3, with a range of 10-25 words. Three of the six Prompt test utterances were single sentences, two were composed of two sentences, and one was composed of three sentences. Prompt test utterances ranged from 12 to 19 words in length, with a mean of 15.7 words.

Half of the test utterances were among the recordings included in the acoustic inventory available for synthesis, while the other half were not included in the inventory. Of the six WSJ test sentences, four were included in the acoustic inventory, while two of the six Prompt test utterances were included in the acoustic inventory. Thus, the evaluation also provided a comparison of the perceived quality of these two cases of test utterances.

All three prosody models were implemented within the Festival Speech Synthesis System [6] and test stimuli were generated using the same voice, duration model, and other modules, for two speech representations, HNM and PSOLA. The use of two different synthesizers in the evaluation increased the generality of our results, since prosodic models can and do interact with the methods with which the resulting speech is synthesized.

## 7.3. Test Conditions

Each of the 12 test utterances were synthesized with each of 12 combinations of synthesis conditions. There were a total of six variations of prosody generation used, and two synthesizers. The segments and durations specified were the same in every case. There were a total of 144 test items. The test conditions examined were as follows:

- Two synthesizers, HNM [9] and PSOLA [6], were used to synthesize the test utterances.
- A natural prosody control case (Nat): A smoothed F0 contour was extracted directly from the recorded test utterances using the "icda" program available with the Festival system. The contour and the segment specification were fed to the synthesizer and a waveform was synthesized.
- A rule-based case (Rule): A contour was predicted based on the segmentation and the ToBI labels for the test utterances.
- A tilt case (Tilt): A contour was predicted from the derived Tilt accents and boundary markers. This data was then passed to the synthesizer module for processing.
- A vector-quantized PaIntE-No ToBI case (VQN): A contour was generated by using accent placement information and markers for rising or falling boundary (as opposed to using both accent location and accent type information), and synthesis used this contour. The number of VQ codebook entries chosen was 8.
- A vector-quantized PaIntE-ToBI case (VQ8): Information about both accent placement and type were used to construct an F0 contour. The number of VQ codebook entries chosen was 8, as for the previous case.
- A second vector-quantized PaIntE-ToBI case (VQ16): The contour was generated as for the previous case, but with a VQ codebook size of 16.

## 7.4. Formal Listening Test Procedure

Test utterances were 16 bit linear speech files with a sampling frequency of 16 kHz and were energy normalized. They were 40-6500 Hz bandpass filtered by a Wavetek Brickwall Filter System 716 and presented to listeners over Sennheiser HD 250 Linear II calibrated headphones.

Subjective ratings of each test utterance were collected from each of 43 listeners, resulting in a total of 6,192 observations. Listeners were native American English speaking adults with no known hearing loss. All listeners were experienced with previous voice quality listening tests, but they were unfamiliar with text-to-speech synthesis. They were tested in four groups of 10-11 listeners each. For each test utterance, listeners were presented a 5-point Mean Opinion Score (MOS) rating scale from which to indicate their judgments using a touch sensitive screen. On the rating scale, 5=excellent, 4=good, 3=fair, 2=poor, and 1=bad. The order of presentation of test utterances was independently randomized for each group of listeners tested.

A brief familiarization session preceded testing, during which listeners were presented samples representing a wide range of the variation among the test utterances presented during the experiment. They were also given practice in using the rating scale and touch screens. The listening test lasted approximately one hour, including the initial instructions and practice session.

## 7.5. Statistical Analysis

A repeated measures Analysis of Variance (ANOVA) with the following three factors was performed: Style(2), Synthesizer(2), Prosody Model(6). A second ANOVA was performed to compare ratings of each test utterance independently. The three factors of the second ANOVA were: Utterance(12), Synthesizer(2), and Prosody Model(6). Statistical significance of main effects and interactions was determined from the resulting F statistics, with $p<0.05$.

# 8. RESULTS

There was a significant main effect of Prosodic Style ($F=22.837$; $df=1,42$; $p<0.0001$). The WSJ test utterances, whose mean rating was 3.37, were rated higher than the Prompt test utterances, with a mean rating of 3.23.

There was no significant difference in overall ratings between the two synthesizers tested, HNM and PSOLA. However, there was a significant interaction ($F=6.201$; $df=1,42$; $p<0.017$) between Prosodic Style and Synthesizer. The relative superiority of the WSJ style over the Prompt style was greater for PSOLA (for which the difference in ratings was 0.17) than for HNM (with a difference of 0.11). PSOLA had higher ratings than HNM for WSJ utterances, and HNM had higher ratings than PSOLA for Prompt utterances.

There was a significant main effect for Prosody Model (F=53.839; df=5,210; p<0.0001), indicating that there were significant differences in ratings among the six models evaluated. The natural prosody control condition received higher ratings than any of the other models. The VQ16 model was rated superior to any of the other remaining models. There was no significant difference among VQ8, Rule, and VQN ratings, and Tilt was lower than any of the other models. Figure 1 shows the mean ratings received by each of the prosody models.

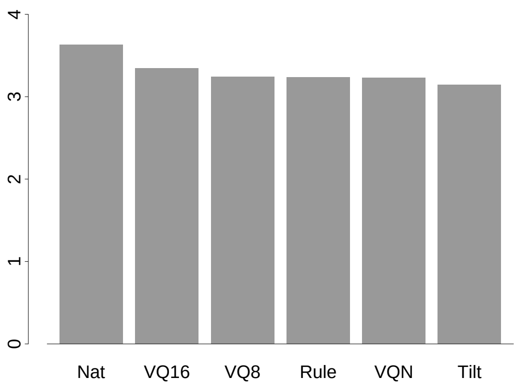

**Figure 1:** Mean Ratings of Prosody Models

There was a significant Prosodic Style by Prosody Model interaction (F=4.498; df=5,210; p<0.001), which primarily reflected large differences in rankings of the VQN and Rule models between the WSJ and Prompt prosodic styles. For the WSJ utterances, Rule was the third-highest rated model, while VQN was ranked with Tilt at the bottom of the rankings. For the Prompt utterances, VQN was the third-ranked model, whereas the Rule model was equivalent to Tilt in last place. It may be that the ToBI training materials on which the Rule model was based are more similar to the WSJ style of prosody than to the Prompts style.

There was a significant Synthesizer by Prosody Model interaction (F=5.389; df=5,210; p<0.0001), which also primarily reflected differences in rankings of the VQN and Rule models depending on the synthesizer. For PSOLA, the VQN model was ranked third, and the Rule model was ranked fifth, of the six models tested. For HNM, however, the Rule model was ranked third, and the VQN model was ranked fifth.

Finally, the ratings of test utterances represented in the acoustic inventory were compared to the ratings of test utterances not included in the acoustic inventory. The mean rating for the six utterances in the acoustic inventory was 3.4577, and the mean rating for the six utterances not in the acoustic inventory was 3.1470. The difference of 0.3107 in mean ratings represents a significant preference for synthesized utterances that were actually in the acoustic inventory. Higher ratings for utterances in the acoustic inventory is attributable to the availability of a full sequence of acoustic units for the synthesis of the utterance. To what extent the sequence of units in the inventory is actually used for synthesis of the utterance depends on how closely the target prosody and phone sequence correspond to those of the recorded natural utterance in the inventory. The greatest difference in ratings between in-inventory versus out-of-inventory utterances would be expected for the natural prosody controls. For these, the mean rating for in-inventory utterances was 3.9419, and that for out-of-inventory utterances was 3.3101, a difference of 0.6318. There is a striking difference between natural prosody Prompts and WSJ test utterances when in- versus out-of-inventory ratings are compared. The mean rating for natural in-inventory Prompts was 4.3663, and that for out-of-inventory Prompts was 3.1396, whereas the natural in-inventory mean for WSJ sentences was 3.7297, and the out-of-inventory mean was 3.6512.

## 9. CONCLUSIONS

In this paper we describe three models for prosody generation. The models were used in a formal listening test in which listeners rated on a 5-point scale the overall speech quality of test utterances.

Conclusions from the experiment are as follows:

- The VQ systems seem to provide a measure of robustness. For the Tilt system the various parameters of the model are predicted independently. For the various VQ versions we are essentially predicting the parameters together as a group. We believe this plays a significant role in producing good F0 contours.
- The lower ratings for the Prompt style can be attributed to three possible factors: (1) there was a smaller database available for Prompts, and there were numerous repetitions of utterances in it. This would have implications for both training and testing; (2) there are fewer instances of Prompt test utterances in the acoustic inventory; and (3) it may be more difficult to model the Prompt material.
- The interaction between prosody model and unit selection inventory means that for a TTS system with a standard fixed diphone inventory, we might expect less overall difference among the various models, although the overall order of the results should remain the same. We are of course interested in maximizing synthesis quality for a given input and the interaction is part of how the data-driven unit selection approach comes into play. The better we can specify prosody that exists in the inventory, the better the resulting synthesis quality.
- Unsurprisingly, models that used more precise input information (ToBI label types as well as locations) were able to generate contours that were more acceptable to listeners than those that only used accent location as input. The Tilt model was the least demanding in terms

of input information, not even requiring precise accent placement information.

This does not imply that models which accept only accent placement information as input are less useful. There is the rather intractable problem of reliable accent generation to be considered in this respect. However in cases where accent placements *and* type information is available, for example in a concept to speech system, it would seem sensible and useful to use it.

## 10. REFERENCES

1. A. Black, P. Taylor, and R. Caley. The festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival.html, 1998.
2. K. Dusterhoff and Black A. Generating $F_0$ contours for speech synthesis using the tilt intonation theory. In *Proc. ESCA Workshop on Intonation*, Athens, Greece., 1997.
3. Matthias Jilka. *Regelbasierte Generierung natürlich klingender Intonationsmuster des Amerikanischen Englisch (Rule-based generation of naturally sounding intonation patterns of American English).* University of Stuttgart, Institute of Natural Language Processing, University of Stuttgart, 1996.
4. Gregor Möhler. *Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese (Theory-based modeling of German intonation for speech synthesis).* University of Stuttgart, Institute of Natural Language Processing, University of Stuttgart, 1998.
5. Gregor Möhler and Alistair Conkie. Parametric modeling of intonation using vector quantization. In *Third International Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.
6. E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 (5/6):453–467, 1990.
7. K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labeling english prosody. *ICSLP*, 2:867–870, 1992.
8. P. Taylor and A. Black. Synthesizing conversational intonation from a linguistically rich input. In *Proc. ESCA Workshop on Speech Synthesis*, pages 175–178, Mohonk, NY, 1994.
9. J. Schroeter Y. Stylianou, T. Dutoit. Diphones concatenation using a harmonic plus noise model of speech. *Proc. EUROSPEECH*, Sept. 1997.