DATA DRIVEN INTONATION MODELLING OF 6 LANGUAGES

Jeska Buhmann 1, Halewijn Vereecken 1, Justin Fackrell 2, Jean-Pierre Martens 1 and Bert Van Coile 1,2

1 ELIS, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
2 Lernout & Hauspie Speech Products NV, Flanders Language Valley 50, B-8900 Ieper, Belgium
E-mail: jeskab@elis.rug.ac.be

ABSTRACT

A method for creating multi-lingual intonation models is described. The method adheres closely to the pioneering work of Traber, in that a recurrent neural network (RNN) predicts a number of $F_0$ values per syllable. An important aspect of the work presented here is the selection of linguistic and prosodic features that are suitable for predicting the observed intonation phenomena in different languages. Another aspect is the use of automatic labelling techniques for the preparation of the training data. Experiments on six languages demonstrate that even though there are differences in performance across languages, it is possible to obtain good results for all six languages. More importantly, making use of automatic labelling techniques for the construction of the training corpora, tends to give better results than making use of manual labelling techniques.

1. INTRODUCTION

Prosody is the feature of TTS systems which is most in need of improvement. An important aspect of prosody is intonation. This contribution describes a method for creating intonation models of isolated sentences. The goal is to predict fundamental frequency ($F_0$) contours, given the orthography. The emphasis is on automatic data-driven techniques. Data-driven models can easily be adapted to different speakers, different text styles (e.g. news, children’s stories, instructions) and different languages. Besides an automatic modelling technique, we also want to adopt automatic labelling techniques during the construction of the training corpora.

An important aspect of this work is the selection of universal (language independent) linguistic factors that are important for predicting observed intonation phenomena. These selected linguistic features (e.g. part-of-speech, type of punctuation) are then combined with prosodic features such as word boundary strength, word prominence and phone duration, which were themselves predicted by prosody models [3]. Indeed, the prosody prediction of our TTS system comprises four successive steps:

1. The prediction of prosodic boundary strength (PBS) from linguistic features. PBS describes the strength of the prosodic break between two words and is measured on an integer scale from 0 to 3.
2. The prediction of word prominence (PRM) from linguistic features and PBS. PRM describes the prominence of a word relative to other words, and is measured on an integer scale from 0 to 9.
3. The prediction of phone duration (DUR) from linguistic features, PBS and PRM.
4. The prediction of $F_0$ contours from linguistic features, PBS, PRM and DUR.

Each model obviously uses its own set of linguistic features derived from the orthography.

A key element in our approach is the fact that the training corpora are constructed using automatic labelling techniques. To begin with, there is an automatic phonetic segmentation and labelling (annotation) tool [15] which takes a phonotyphical phonetic transcription as an input. The latter is obtained via the grapheme-to-phoneme component of our TTS system. Next, there is an automatic PBS and PRM labelling [14]. Obviously, some manual labelling was involved in the creation of training material for these automatic labelling tools. To train the automatic phonetic annotation, five minutes of manually annotated speech seem to suffice [15]. To train the automatic PBS/PRM labeller, we need forty minutes of manually labelled speech. Three things need to be emphasized:

1. The manual labelling of PBS and PRM is based on perception only, and is performed by native listeners. The manual labeller is not instructed to pay attention to particular acoustic cues. E.g. a PRM value of 5 can at one time be attributed to segmental lengthening, while at another time it can result from an $F_0$ excursion.
2. The automatic phonetic annotation can automatically be adapted from one speaker to another and, more importantly, from one language to another [15].
3. The automatic PBS/PRM labeller can not automatically be adapted from one speaker to another, nor from one language to another. Different speakers may use different strategies for signalling boundaries and accents to the listener. Moreover, these strategies may differ significantly across languages.

The intonation model we propose here adheres closely to the pioneering work of Christof Traber [12]: a recurrent neural network (RNN) is used to predict a number of pitch values per syllable, given a set of orthography-derived features for that syllable. Our work differs from Traber’s in that we use automatically labelled databases, that we tested different feature sets, and that we investigated the importance of these features in a multi-lingual and multi-speaker experiment involving 6 languages and 6 different speakers.
2. TAXONOMY OF EXISTING SYSTEMS
Two major distinctions can be made in the way intonation has
been modelled so far [4]:

2.1. Phonological Versus Phonetic Models
A phonological model uses a phonological representation of \( F_0 \).
Such a representation is descriptive and discrete. It uses an in-
ventory of abstract phonological categories, with each category
having its own linguistic function. The most notorious example
is perhaps the tonal tier of the ToBI labelling system [1]. ToBI
specifies an inventory of tones: one set is used to mark accented
syllables, while another set is used to mark phrase boundaries.
Each tone marks a different type of accent or boundary.

A phonetic model is developed from acoustic (\( F_0 \)) data. It at-
ttempts to describe \( F_0 \) movements and it is usually continuous in
nature. Often, the description of \( F_0 \) movements is linked in some
way to the linguistic level. The Tilt intonation model [11], for ex-
ample, describes pitch accents and boundary tones via rising and
falling quadratic functions. A pitch accent can be composed of a
rising function, a falling function, or a rising followed by a falling
function. Speech parts between intonational events are described
by straight line interpolations. The amplitudes and durations of the
rising and falling functions, combined with the position of the
pitch accent/boundary tone in the \((F_0, f_0)\) plane, together constitute
the basis for the Tilt description of \( F_0 \) contours (5 continuously-
valued parameters per event). The Fujisaki model [7] views an
\( F_0 \) contour as the sum of a base \( F_0 \) value, phrase components and
accent components (in the log \( F_0 \)-domain). Phrase and accent
components are generated by respectively passing impulse and
step functions through second order filters. The timings and am-
plitudes of the impulse and step functions constitute the phonetic
representation of \( F_0 \).

Traber’s representation of \( F_0 \) [12] merely consists of samples of a
smoothed and interpolated \( F_0 \) contour. It does keep pace with the
syllabic structure of the utterance, but it has no other links to the
linguistic level.

2.2. Sequential Versus Superpositional Models
It is commonly acknowledged that an \( F_0 \) contour is the result of
many interacting factors, each having a different temporal scope
(phone, syllable, word, phrase, sentence, paragraph). A superpo-
sitional model attempts to model some or all of these factors sep-
ately, and combines the partial models to an \( F_0 \) contour. The
Fujisaki model is such a model. Sequential models directly gen-
erate \( F_0 \) values from left to right as a sequence of \( F_0 \) values or
movements. The Tilt model and Traber’s model fall in this group,
along with many ToBI-based models.

2.3. Basic Considerations for Our Model
There are a number of shortcomings to phonological models, at
least from our point of view. Phonological models are not eas-
ily ported from one language to another: if at all possible, the
inventory of categories must be thoroughly reviewed by expert
linguists. Keller et al [6] argue that a phonology such as ToBI is
for instance very suspect for French \( F_0 \) contours. Each phonolog-
cal representation of intonation relies on a linguistic theory, and a
general theory of intonation which is acknowledged by everyone,
is very much absent. We therefore figured that a phonetic model
is much more useful in a multi-lingual context.

The choice between a sequential and a superpositional approach
is not critical within the framework of our research.

The intonation models presented in this paper use a RNN to pre-
dict 5 pitch values per syllable from orthography-derived features
such as part-of-speech, punctuation type, PBS, PRM, etc. Our
models can thus be termed sequential and phonetic.

3. THE INTONATION MODEL

3.1. Structure of the RNN
The RNN has about 80 inputs; the exact number depends on
the language because categorical features do not always have the
same number of categories in all languages. The first and second
hidden layer comprise 20 and 10 hidden units respectively. There
are 5 outputs, corresponding to the 5 \( F_0 \) values. Each layer is
fully forward connected to the next layer. Furthermore, each unit
on the first hidden layer has a recurrent connection to a context
unit which is added to the input layer. The RNN operates from
left to right, starting from the first syllable in the sentence.
Each RNN was trained and tested using the Stuttgart Neural Net-
work Simulator [16], which is freely available on the Internet.
All networks are of the Elman type. Training was done with the
backpropagation-through-time algorithm. The learning rate
was initialised to 0.2, but gradually decreased during the training.
The training continued until maximum performance on a cross-
validation set was achieved.

3.2. Inputs of the RNN
There is usually no problem in assigning word- or even sentence-
level features to a syllable (e.g. part-of-speech, sentence type). It
does however pose a problem to include prosodic features such as
PBS and PRM, which are actually word-level features. It does not
take too much imagination to see that each syllable within a word
can not simply inherit the PBS or PRM of that word. For exam-
ple, if the word is followed by a major break, it would be foolish
to give each syllable boundary the same high PBS value. The
same goes for PRM: not all syllables within a word are equally
important. Therefore we have adopted the following rules:

- A word-final syllable inherits the PBS of the word. The other
  ones get a PBS=0.
- A primary stressed syllable inherits the PRM of the word, a
  secondary stressed syllable is assigned half the word PRM,
  and any other syllable gets 10% of the PRM.

A syllable boundary having a PBS \( \geq 2 \) is considered a phrase
boundary. A syllable having a PRM \( \geq 5 \) is considered an ac-
ccented syllable. This consensus was the result of listening tests.
We would like to recall though, that a PRM value of 5 does not
necessarily implicate a pitch accent: the perception of this value
may well have been triggered by segmental lengthening alone.

Eventually, a set of 25 different features was conceived to de-
scribe a syllable. Among them you will find PBS, PRM, number
of words in the sentence, position of the word in the sentence,
word frequency, part-of-speech, accentability, position of the syl-
vable with respect to the main accent in the phrase, distance (ex-
pressed as a number of syllables) to preceding and succeeding ac-
cents and punctuation symbols, punctuation type, duration of the
pause following the current word, duration of the syllable onset
and rhyme (normalised with respect to the mean and deviation),
position of the syllable with respect to the beginning and end of the word, etc.

For PBS and PRM, we also used contextual information, namely values of the preceding and succeeding syllable. All other features refer to the current syllable only. Categorical features (e.g. part-of-speech) are represented via 1-of-N coding. For Dutch, for example, the 25 features thus amount to 80 RNN inputs.

### 3.3. RNN Outputs

The 5 RNN outputs correspond to 5 equidistantly spaced $F_0$ values of the investigated syllable. The training targets are obtained as follows:

- The $F_0$ values (in Hz) were obtained via AMPEX [13] and transformed to the semitone domain, using the average $F_0$ of the speaker (E[$F_0$]) as a reference.
- The $F_0$ samples are then subjected to a smoothing and interpolation process, similar to (yet different from) the one described by Taylor [10]:
  1. Voiced frames having an $F_0$ outside a broad range around the reference (E[$F_0$]) are considered unvoiced.
  2. The remaining voiced frames were subjected to a 9-point median filter operator. Unvoiced frames within the 9-point window are ignored for computing the median. E.g. if there are 4 voiced frames preceding the current voiced frame, and 4 unvoiced frames following, the median will be computed over the 5 voiced $F_0$ values alone.
  3. Unvoiced portions of the $F_0$ contour are filled using linear interpolation.
  4. The (now fully voiced) contour is additionally smoothed using a $\cos^2$-filter.

During testing, frame level pitch predictions are constructed using linear interpolation of the estimated syllable level values.

### 4. EXPERIMENTAL EVALUATION

#### 4.1. Speech Databases

Our $F_0$ modelling approach was evaluated on 6 databases corresponding to 6 different speakers and 6 different languages [5]: American English, Dutch, French, German, Italian and Spanish. Each database contains about 1200 isolated sentences, representing about 120 minutes of speech. The sentences include a variety of text styles, syntax patterns and sentence lengths. The majority of the sentences is declarative though. The recordings were made with professional native female speakers (one speaker per language). All databases are carefully hand-labelled with respect to PBS and PRM. Each database is partitioned into a training set (75%), a cross-validation set (10%) and a test set (15%).

#### 4.2. Results

The test set results mentioned in this section are obtained using labelled instead of predicted prosodic features. I.e. the features PBS, PRM and DUR are obtained by taking the speech signal into account. This allows for an evaluation of the $F_0$-modelling part of prosody alone, and it complies with the mainstream approach adopted in the intonation modelling literature. Obviously, in a real TTS system, the predictions of PBS, PRM and DUR by their respective models will have to be used. We normally use automatically labelled prosodic features, but for comparison, we also tabulate some $F_0$ modelling results obtained using manually labelled PBS and PRM. In practice, 40 minutes of labelled speech would have been sufficient, but as we wanted to compare $F_0$ modelling results using automatic versus manual labels, we took the full database (120 minutes) into account for the manual labelling case as well. Unfortunately, we could not dispose of any hand-marked DUR.

The $F_0$ modelling performance is measured by calculating on each test set the Pearson correlation coefficient ($r$) between the smoothed observed contours and the predicted pitch contours, the root mean square error ($e$) and the percentage of explained variance (EVAR). Only those frames where the original observed contour was voiced, were taken into account.

The test set results are given in table 1. In order to interpret them one has to bare in mind that the higher the speaker’s $E[F_0]$ is, the larger the deviations with regard to $E[F_0]$ are bound to be. It is hence logical that $r$, RMSE and EVAR get worse if $E[F_0]$ gets larger. We therefore introduced two additional error measures:

1. Let $V_e$ be the variance of the observed, and $V_p$ the variance of the predicted $F_0$ values then the measure $E_x = \sqrt{V_e/V_p}$, describes the amount of excursation that was modelled with respect to the amount of excursation in the original contour.
2. The coefficient of variance $C_v = e/E[F_0]$ shows the amount of discrepancy between observed and predicted contours relative to $E[F_0]$.

According to the Pearson correlation coefficient $r$ and the percentage of explained variance EVAR, both Italian and Spanish seem to be modelled better than the other four languages (with the

<table>
<thead>
<tr>
<th>Language</th>
<th>Automatic PBS &amp; PRM</th>
<th>Manual PBS &amp; PRM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E[F_0]$</td>
<td>$e$</td>
</tr>
<tr>
<td>Dutch</td>
<td>171.0</td>
<td>0.71</td>
</tr>
<tr>
<td>English</td>
<td>216.4</td>
<td>0.71</td>
</tr>
<tr>
<td>French</td>
<td>212.3</td>
<td>0.58</td>
</tr>
<tr>
<td>German</td>
<td>222.0</td>
<td>0.58</td>
</tr>
<tr>
<td>Italian</td>
<td>195.2</td>
<td>0.71</td>
</tr>
<tr>
<td>Spanish</td>
<td>216.3</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 1. Intonation modelling performance for 6 languages with (1) automatic prosodic labels and (2) manual prosodic labels: ($E[F_0]$) the speakers mean $F_0$, ($e$) RMSE (in Hz), ($r$) Pearson’s correlation, ($EVAR$) the percentage of explained variance, ($Ex$) the excursion of the pitch contour which is modelled and ($C_v$) the coefficient of variance.
worst results for German). The coefficient of variance \( C_v \) seems to indicate that the intonation model works best for Spanish, and produces slightly worse results for the other languages.

The measure \( Er \) demonstrates that the modelled excursion of the intonation contour comprises about 70% of the excursions in the observed intonation contours. This indicates that the model predicts pitch values in a rather neutral way: it most probably will never predict non-standard intonation movements or very high or low pitch excursions.

Altogether it seems that the results for Dutch, English and French are rather similar, those for Italian and Spanish are somewhat better, and those for German are worse. In any case, the conclusion may be drawn that our intonation model is capable of producing seemingly acceptable results for different languages.

The table also demonstrates that the use of automatic prosodic labels yields better results than the use of manual prosodic labels. This is most likely due to the fact that an automatic labeller is more consistent than a human labeller.

### 4.3. Comparison with other Research Results

As most published results for English and for female speakers were obtained by training on a smaller database than ours, we have performed an additional experiment using just 45 minutes of training speech. The results of our two experiments are compared to previously reported data in table 2.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Training size (min)</th>
<th>( e )</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>this paper</td>
<td>120</td>
<td>29.4</td>
<td>0.66</td>
</tr>
<tr>
<td>this paper</td>
<td>45</td>
<td>30.8</td>
<td>0.62</td>
</tr>
<tr>
<td>Möhler [8]</td>
<td>104</td>
<td>39.9</td>
<td>-</td>
</tr>
<tr>
<td>Dusterhoff [2]</td>
<td>45</td>
<td>32.5</td>
<td>0.60</td>
</tr>
<tr>
<td>Ross [9]</td>
<td>48</td>
<td>34.7</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2. Our results (in Hz) versus other results from literature.

The data indicate that our approach is sound, and that it yields better models when more training data is available.

### 5. CONCLUSION

In this paper a data driven method for intonation modelling of different languages is described. A set of 25 features is used to predict 5 pitch values per syllable by means of a RNN. These pitch values are transformed into pitch contours by means of interpolation. From the results it can be concluded that our intonation model works well for the 6 different languages that were studied.

### 6. ACKNOWLEDGEMENTS

This research was performed with support of the Flemish Institute for the Promotion of the Scientific and Technological Research in the Industry (contract IWT/ADV/980277).

### REFERENCES


A method for creating multi-lingual intonation models is described. The method adheres closely to the pioneering work of Traber, in that a recurrent neural network (RNN) predicts a number of $F_0$ values per syllable. An important aspect of the work presented here is the selection of linguistic and prosodic features that are suitable for predicting the observed intonation phenomena in different languages. Another aspect is the use of automatic labelling techniques for the preparation of the training data. Experiments on six languages demonstrate that even though there are differences in performance across languages, it is possible to obtain good results for all six languages. More importantly, making use of automatic labelling techniques for the construction of the training corpora, tends to give better results than making use of manual labelling techniques.