Modelling Pitch Accent Types for Polish Speech Synthesis

Dominika Oliver[1] *and Robert A. J. Clark*[2]

[1] Institute of Phonetics, Saarland University, Saarbrücken, Germany
[2] CSTR, University of Edinburgh, Edinburgh, UK
dominika@coli.uni-sb.de, robert@cstr.ed.ac.uk

Abstract

We describe a Polish prosody modelling module for the Festival speech synthesis system. The module uses classification and regression trees for accent type prediction and a linear regression technique for F0 contour generation for these contours. The techniques used to attempt to overcome problems with the only available data are shown. We demonstrate how improvements were achieved by the use of a modified F0 stylisation, accent type clustering and language specific features. Results of a formal perception study show a significant preference for the new intonation model over the original one.

1. Introduction

In order to get high quality speech synthesis for Polish, a good prosodic model needs to be developed. Modelling prosody for diphone based concatenative speech synthesis involves generating pitch contours based on a linguistic description of a language. The training of a model requires a viable linguistic description of the intonation of the language and a suitably annotated speech database.

We outline the characteristics of Polish prosody and describe the linguistic resources available in Polish. An automatic prosodic re-annotation based on a modified F0 stylisation and clustering of accent types is presented. We then discuss two models we have built for use with the Polish voice [1], a CART [2] model for accent prediction and a Linear Regression [3] model for contour generation. We pay particular attention to the language specific details of these models and the techniques used to overcome problems with the resources that are available in Polish.

2. Background resources

In order to successfully develop a prosodic model, we need three components: a language specific description of intonation, an annotated speech corpus on which the model can be trained and a system in which the model can be implemented and tested. This section looks closely into these three components.

2.1. Intonation description

The analysis and description of Polish intonation has a long history with the most notable analysis by Jassem [4]. Jassem's theoretical model, based on the British school, assumes the existence of three tone heights, low (L), mid (M), and high (H). More recent studies, e.g. [5], using modern computation methods and new speech corpora, have provided a statistical analysis of this model. Its description consists of five distinct tone heights, marked by xL, L, M, H, xH ranging from extra low to extra high respectively and define acoustic parameters for two prenuclear (H, L) and nine nuclear accents (HL, LH, LM, ML, HM, MM, MH, xL, LHL).

As this analysis was carried out with speech technology applications in mind, we take it as a reliable starting point for the prosodic modelling undertaken in this study.

2.2. Speech Resources

Having an intonation description system for a language is one thing, having a suitably annotated speech database is another. In fact, until recently there was no such prosodically annotated database for Polish, which explains the lack of previous work on generating Polish intonation using data driven methods.

The current study uses the PoInt speech database of Polish, which has been recorded as part of the Polish Intonation Database Project [6]. This database includes recordings from 43 speakers of mixed sex and contains a variety of discourse types: fragments of read literary texts and quasi-spontaneous monologues as well as map task based dialogues. PoInt is the first database for Polish which has been annotated with prosodic information based on Jassem's revised description, marking intonation tone heights with labels xL, L, M, H, xH [6]. The accented syllables contain one letter labels and if the accented word is followed by an IP boundary the post-accented syllable also has a tone height marker followed by a boundary marker '|'. Such annotation gives us quantitative information on the contour shapes through tone heights which have been impressionistically labelled by trained human labellers.

2.3. Polish TTS system

The system within which the current work is carried out is a Polish voice module for Festival [7]. Festival is a general multilingual speech synthesis system developed at the Centre for Speech Technology Research at the University of Edinburgh. The Polish module uses concatenative diphone based synthesis.

The work described here improves upon the very simple intonation models that this module relies upon. The original voice uses a simple prosody prediction and generation module which puts an accent on content words by realising a hat shaped pattern on the stressed syllables of each content word (see Fig.1).

3. Preprocessing

As described in Section 2 the PoInt database at our disposal has been prosodically annotated with tone heights. This annotation needed to be transformed to accent type annotation in order to implement it into our accent prediction modules.

A manual prosodic annotation is considered to be a difficult task and can result in a low inter-annotator agreement. For the database used no such figures were available. Instead we

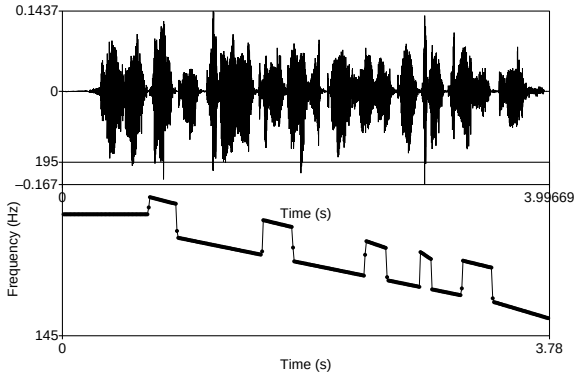

Figure 1: *Hat pattern F0 contour for a sentence "Florentynka przenikliwie spojrzała w oczy Kłoski".*

derived an acoustic description for all prosodic labels in the database.

We decided to minimise annotator bias by using an automatic procedure [8] to prosodically relabel the database based on the original annotation. This procedure consisted of first applying a modified F0 stylisation method to derive a continuous F0 curve for all sentences in the database and then using unsupervised clustering to derive accent classes.

3.1. F0 stylisation

The F0 curve was stylised by the Momel algorithm [9], modified in boundary conditions. In order to span the whole signal we add, if present, original F0 points to the Momel target points at the boundary locations. In cases of unvoiced material at both ends of the signal, the values before the F0 onset and after the F0 offset are replaced, as in the preprocessing stage for Fujisaki model parameters extraction, with the first and last F0 value, respectively, found in the extracted contour. In this way elements of the stylised curve are enriched with the initial and final variations of F0 essential for capturing boundary prosodic events. The new continuous F0 curve is then derived by quadratic interpolation of the Momel target points and the original F0 points added at the beginning and end of an utterance.

3.2. Clustering accent types

In this study, a two stage procedure was used to cluster accent types present in the PoInt database. First, the stylised F0 contours of PoInt database sentences were used as input to Self-Organising Maps (SOMs) [10], a vector quantisation method which performs clustering of instances in the form of acoustic feature vectors. The method determines the prototypical instance and then adjusts its coordinates so that it becomes even more similar to the training data. As a result, we obtained similarity clusters that can be seen as soft edged classes or fuzzy sets emerging from statistical correlations. To derive the final clusters from the resulting topological representation of the original data, hierarchical agglomerative clustering was applied.

The three resulting clusters had to be analysed in order to determine their correspondence to the sentence linguistic structure. The contours in the first group exhibited the steepest rising and falling movement, with F0 peak positioned in the middle of the syllable. These contours reached the highest F0 maximum values and the lowest F0 minimum values of all clusters. Members in this group were mostly associated with emphatic focus or a wh-word in a question. The second group was characterised by a very early F0 peak, mostly starting at the syllable start or early in the onset forming a plateau, followed by a gradual fall ending at the speaker's mid or bottom range. 90% of the pitch accents in this cluster were phrase final and characteristic of a final fall. The pitch accents in the third group exhibited a very late F0 peak, either in the syllable offset or on the following syllable. This type of contour is found mostly in phrase final position and is typically associated with either a continuation rise or a yes/no question. The resulting cluster groups 2 and 3 correspond to the accent groups identified in connected speech by Demenko [5], where the intonation contours were conflated into rising (R), and falling (F). Additionally, contours in our first cluster can be seen as rising-falling (LHL) associated with emphasis and which, in view of low occurrence in speech material in [5], were conflated. In our study, LHL was classified as a separate group. Consequently, the PoInt corpus was relabelled according to the new pitch accent type membership (LHL, R, F), and the new labels were used as input to the accent prediction module described in Section 4.

4. Accent prediction

Classification and regression trees (CARTs) are often used in speech synthesis for duration modelling, but they can also be applied to accent prediction [11]. Trees are constructed by a data driven training process. A tree is derived by a set of yes/no or if/then questions relating to the data in order to predict the dependant variable. As a classifier, the tree partitions the training data into classes so that the class with the highest probability is chosen to classify an unknown case. As a regression tree, the tree is designed to minimise the error in the predicted variable.

The main advantage of this method is that it can deal with non-linear data in a reasonable way. However, the training procedure can get stuck at local maxima. This can be overcome to an extent by providing the model with features that are thought to most likely to influence the placement of pitch events.

In the original implementation for Polish accent prediction there was no support for different accent types. In contrast, for languages like English, models built from databases with labels based on ToBI [12] or similar descriptions have been implemented. In our case, this paradigm has been modified to use the PoInt annotation after relabelling described in Section 3 based on a set comprising of LHL, R(ising), and F(alling) accents. The spontaneous speech part of PoInt was excluded because of problems relating to speaker overlap.

For accent type and place prediction we initially used the following features: position of syllable in phrase/word, strength of break, stress, number of (stressed, unstressed, accented) syllables since/till last phrase break, number of syllables since/till last accented syllable, number of minor phrase breaks since last major phrase break.

We then introduced language specific features to improve accent type prediction, namely, additional information regarding the position of a syllable within a word. It is motivated by the fact that in Polish the last content word usually receives the phrase accent and stress occurs generally on the penultimate syllable of a word. To accommodate this the following features were provided:

- the word is final in a phrase
- the syllable is penultimate in a word
- a simplified part of speech (content/function word)

Table 1 shows the confusion matrix for accent type assignment in the final model including category NONE corresponding to unaccented tokens. A 7% improvement was noted as a result of including the language specific features.

Table 1: *Accent prediction confusion matrix.*

Accents	LHL	F	R	NONE	Accuracy
LHL	673	59	91	2	81.6%
F	92	725	146	2	75.1%
R	22	129	774	0	83.7%
NONE	16	24	14	797	93.7%
	Correct 2969/3566			Total Accuracy 83.3%	

The features that had the highest correlation were the number of syllables since (0.9045) and till a phrase break (0.9034), position in word (0.9065), number of accented syllables before next phrase break (0.9086), strength of the break after next syllable (0.9107) and stress on the syllable (0.9117).

5. F0 contour generation

Contour generation in Festival is traditionally carried out using three Linear regression (LR) models [3]. Linear regression models assume that a predicted variable (p) can be modelled as the sum of a set of weighted real-valued factors.

$$p = w_0 + w_1 f_1 + w_2 f_2 + w_3 f_3 + ... + w_n f_n \qquad (1)$$

The factors (f_i) represent parametrised properties of the data, and the weights (w_i) are trained, usually using a stepwise least squares technique. Each of the three models predicts the f0 at a different point of a syllable (start, middle and end respectively). The factors incorporate information like the type of accent present, the position of phrase breaks, syllable stress and syllable position within the text. Each model considers this information for a five syllable window centred on the current syllable. This allows the pitch on syllables around an accented syllable to be affected by the presence of the accent, so that pitch movement is not restricted to occur on the syllable that is marked with the pitch event. For example the peak of an L+H* could occur in the syllable following the one the accent is assigned to.

Based on the accents predicted in Section 4 we built linear regression models to predict f0 at the start, mid point and end of syllables. These replaced the original rule based model which imposed the hat-shaped accents onto a declining baseline.

The advantage of using Linear Regression is that relatively little knowledge about the intonation system in question needs to be known (although arguably this is needed to complete the accent prediction stage). However, we show that the inclusion of language specific features can lead to improvements in the resulting contour.

The training of these models generally requires at least an hour of good quality speech from a single speaker. Our database did not contain enough data per single speaker to build a model from so we decided to investigate the use of multi-speaker data. To be able to use PoInt multi-speaker data, an F0 normalisation procedure was necessary. We adapted the procedure used by [13] to normalise across Tone Groups to normalise across speakers.

The normalisation was carried out using:

$$F0_n = \frac{F0 - \mu_i}{\sigma_i} \qquad (2)$$

where μ_i is F0 mean and σ_i is the F0 standard deviation of the utterance in question

The rescaling uses mean and standard deviation of the database :

$$F0 = F0_n \sigma_D + \mu_D \qquad (3)$$

where μ_D is F0 mean of the database and σ_D the F0 standard deviation of the database.

LR Model 1 is a model which is trained on a set of features that a default Festival English model is trained on except that the features relating to pitch accents have been changed to accommodate the relabelled PoInt prosodic annotation scheme. Fig.2 shows an example waveform and F0 contour synthesised using the new LR Model.

Table 2 shows the RMSE and correlation values for the new LR model for the training and test data, computed for the whole signal.

Table 2: *RMSE (Hz) and correlation for LR Model 1.*

	LR train		LR test	
Position	RMSE	Correlation	RMSE	Correlation
start	42.51	0.72	41.85	0.75
mid	43.5	0.66	43.63	0.68
end	40.3	0.67	40.19	0.7

The results are based on the normalised F0 data from the female and male speakers, which consisted of 865 utterances. The fact that RMSE and correlation figures are not as good as expected is due to the great variability between the speakers and the fact that the normalisation method is not that sophisticated.

6. Perception study

To get a subjective evaluation of the resulting model for F0 contour generation and to measure its relation to the original model a perception study was carried out. To perform the experiment, 15 sentences were selected from the database used to train the LR model. The stimuli were created using Festival TTS system (creation of F0 track) and Praat (pitch replacement). In order to imitate the intonation contour of the original database sentences, APML [14], an XML-based mark-up language, was used to generate the same accent placement and type as marked in the automatically relabelled annotation. Both sets of stimuli varied only in the intonation model applied (original vs. new) keeping other acoustic characteristics intact. Ten native Polish speakers who took part in the study were asked to choose a realisation of the sentence they preferred paying attention to its melody. The experiment was conducted via internet using web forms and the participants had the possibility to listen to stimuli more than once.

The analysis of the results showed that the new intonation model was significantly preferred with a mean 78% preference over the original model ($p<0.001$).

7. Conclusions

The new prosodic models are much better than the original baseline model, given the nature and size of the corpus. Multiple speakers in particular made it difficult to produce a constant dataset from which to train models. Nevertheless, it proves both that Festival can be successfully used as a tool for prosody research for languages other than English or German and that

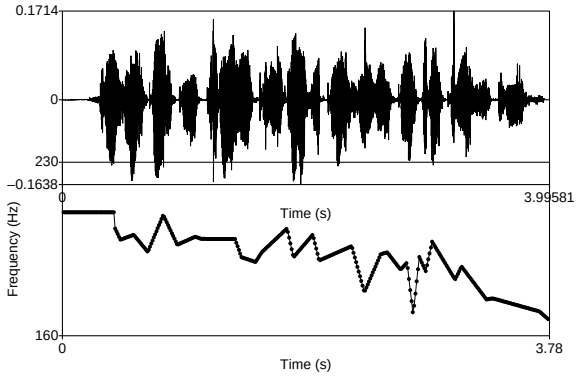

Figure 2: *LR predicted F0 contour for a sentence "Florentynka przenikliwie spojrzała w oczy Kłoski".*

within the limits of the resources, reasonable improvements can be achieved.

Automatic relabelling of the database based on F0 stylised contour resulted in consistent and linguistically meaningful accent type labels. Adding language specific features improved accent prediction, demonstrating the importance of prior linguistic knowledge of the intonation of the language being modelled.

The normalisation of the pitch contours of the different speakers to the same pitch range resulted in limited success. Further work is being carried out to investigate other transforms that are more appropriate than the normalisation technique applied here, for example, the piecewise linear transform used by [15].

We will attempt to further improve the model by making additional recordings containing more tokens of the underrepresented accent types. We will also consider manual alteration of the linear regression parameters [13] to improve the realisation of accents of which there is insufficient data to model properly.

In this work both objective and subjective evaluation has been used to assess the new prosody model on single sentences. The obvious extension will be to incorporate a test to evaluate the acceptance of the new model in a paragraph or dialogue setting.

8. Acknowledgements

The first author is funded by the International Postgraduate College "Language Technology and Cognitive Systems". We would like to thank Maciej Karpiński for making the PoInt Database available for research purposes.

9. References

[1] Oliver, D. Polish Text-to-speech Synthesis. MSc Thesis. University of Edinburgh, 1998.

[2] Breiman, L., Friedman, J. H., Olhsen, J. A. and Stone, C. J. Classification and Regression Trees, Belmont, CA. Wadsworth, 1984.

[3] Black, A., and Hunt, A. "Generating f0 contours from ToBI labels using linear regression", in ICSLP 96, Philadelphia, Pen, 1996.

[4] Jassem, W. Akcent języka polskiego (Accent in Polish). Prace Językoznawcze 31, Komitet Językoznawstwa, Kraków, PAN 1961.

[5] Demenko, G. Analiza cech suprasegmentalnych języka polskiego na potrzeby technologii mowy. Poznań, UAM, 1999.

[6] Karpiński, M., and Kleśta J. "The Project of an intonational database for the Polish language", in Prosody 2000: speech recognition and synthesis. Poznań: Adam Mickiewicz University, 2001.

[7] Black, A., Taylor, P., Caley, R. "The Festival speech synthesis system," <http://festvox.org/festival/>, 1998.

[8] Oliver, D. "Deriving pitch accent classes using automatic F0 stylisation and unsupervised clustering techniques". In Proceedings of Second Baltic Conference on Human Language Technologies, Tallinn, Estonia, 4-6 April, pp. 161–166, 2005.

[9] Hirst, D. J., and Espesser, R. Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix, volume 15, pp. 75–85, 1993.

[10] Kohonen, T. Self–Organizing Maps. Springer Series in Information Sciences, Berlin, Heidelberg, Springer-Verlag, vol. 30, 1995.

[11] Dusterhoff, K., Black, A. and Taylor, P. Using decision trees within the Tilt intonation model to predict f0 contours, in Eurospeech, 1999.

[12] Pierrehumbert, J. B. 1980. "The Phonology and Phonetics of English Intonation." Dissertation, MIT.

[13] Clark, R.A.J. Generating Synthetic Pitch Contours Using Prosodic Structure. PhD Thesis. University of Edinburgh, 2003.

[14] Carolis, B. de, Pelachaud, C., Poggi, I., and Steedman, M. "Apml, a mark-up language for believable behaviour generation," in Life-like Characters, Tools, Affective Functions and Applications, H. Prendinger, Ed., pp.65–85. Springer, Berlin, 2004.

[15] Patterson, D. A Linguistic Approach to Pitch Range Modelling. PhD Thesis. University of Edinburgh, 2000.