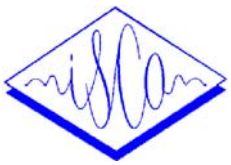

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MULTILINGUAL PROSODY MODELLING USING CASCADES OF REGRESSION TREES AND NEURAL NETWORKS

J.W.A. Fackrell[1], H. Vereecken[2], J.-P. Martens[2], B. Van Coile[1,2]
[1] *Lernout & Hauspie Speech Products NV, Flanders Language Valley 50, B-8900 Ieper, BELGIUM*
[2] *ELIS, University of Gent, Sint-Pietersnieuwstraat 41, B-9000 Gent, BELGIUM*
Email : justin.fackrell@lhs.be

ABSTRACT

This paper describes the use of automatically-trained models (Regression Trees and Multilayer Perceptrons) to predict three prosodic variables – phrase-boundary strength, word prominence and phoneme duration. The models are arranged in a cascade so that the predictions of phrase-boundaries are used as input features to the prominence model, and so on. Cascade models of this type have been constructed for 6 languages, using specially constructed databases, and objective performance statistics are described. For two languages (American English and Dutch) the results of a subjective evaluation experiment suggest that these prosodic models are at least as good as hand-crafted models, and sometimes better. Furthermore, preparing the training data automatically, rather than by manual labelling, seems to have no negative impact on the model performance.

Keywords : Speech Synthesis, Prosody, CART, Neural Networks.

1. INTRODUCTION

Improvement in prosody prediction remains a challenge for producing *really natural* text-to-speech synthesis. The problem of producing good prosody models can be tackled either by using linguistic expertise to hand-craft the models, or by making use of large speech corpora and automatic learning techniques to derive the models automatically. This second approach has attracted much interest, since it offers the potential of rapid model development, and language independence.

Much of the previous work in this field has focused on using *one* technique (MLPs[1], CARTs[2], Sums-of-Products-Models[3]) for the prediction of *one* particular prosodic parameter (phone duration[1,2,3], word prominence [4]), usually for just *one* language. To our knowledge, there have been few attempts to broaden these investigations to span more techniques, more prosodic parameters and more languages ([3,4] are notable exceptions). Our long-term goal is to conceive a language-independent methodology for prosody modelling which can be used to quickly make prosodic models for new languages. The work reported here thus attempts to use a standardized approach of feature extraction and modelling for 6 languages: Dutch, English (US), French, German, Italian and Spanish.

In our approach, prosody is parameterized by four variables: prosodic boundary strength, word-based prominence, phone duration and intonation. The current study includes the first three of these. The models are arranged in a cascade architecture, with the predictions of high-level models being used as input features for lower-level models. Consequently, each "cascade" contains three prosodic models, each of which is either a Multilayer Perceptron (MLP) or a Regression Tree (RT).

The prosody modelling described in this paper follows on from previous investigations into prosodic database design [5], automatic prosodic labelling [6,7] and the reliability of manual prosodic labelling [8].

The paper is organized as follows: Section 2 describes the architecture of the prosody generation module, and describes the MLP and RT modelling strategies. Section 3 describes the 6 specially-constructed databases with which the models were trained. Section 4 presents numerical results on unseen test data, and perceptual evaluation results for two languages, and Section 5 presents conclusions.

2. PROSODY MODELS

Figure 1 shows the architecture of the prosody prediction module. Pitch (INT) and duration (DUR) are determined by a multi-stage process of prediction involving the intermediate variables of prosodic boundary strength (PBS) and word prominence (PRM). The output of each prediction stage is used, together with additional information derived from the text, as the input to the next stage. PBS takes values 0,..,3, PRM takes values 0,..,9 (the choice of scales for these variables was made after experiments into the reliability and consistency of manual labelling [8]) and DUR is measured in *ms*. INT prediction is not covered in this paper.

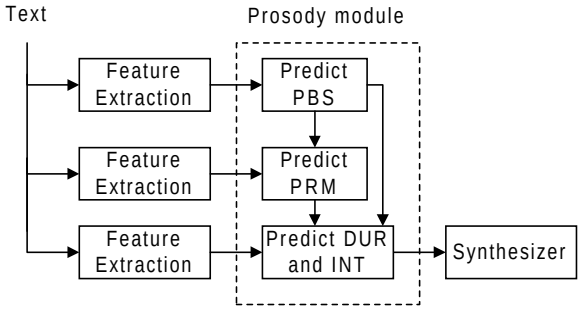

Figure 1 Cascade architecture of prosody prediction.

2.1 Feature Extraction

The key features used to predict each of the three prosodic variables are listed in Table 1. The DUR models use features at all temporal levels from the phone level up to the phrase level, but the PBS and PRM models only use features at the word-level and higher.

Variable	Key features (non-exhaustive list)
PBS	Punctuation type, part-of-speech (POS), distance to prev/next punctuation, sentence length (in words), position in phrase/sentence, word length (in syllables).
PRM	PBS, punctuation type, POS, distance to prev/next punc., syntactic phrase type, sentence length, position in phrase/sentence, word length (in syl.).
DUR	PBS, PRM, Phoneme ID and class (manner, voicing, height, place), phonetic context, consonant cluster size, position in word, lexical stress.

Table 1 Key features for each of the target variables PBS, PRM and DUR. These lists are not exhaustive.

2.2 Modelling

The MLP models used for each of the three prosody components are two-layer perceptrons. The RT models are trained using the techniques described in [9]. For PBS and PRM, both MLPs and RTs predict integer values. For duration, the models operate in the log-*ms* domain, to accommodate the effects of the positive skew of duration distributions [10]. The training methodology, like the prediction architecture, is cascaded, so the PBS model is trained first. Its prediction ("predicted PBS") is added to the PRM feature set. A model of PRM is then trained. The predictions of both PBS and PRM models are added to the DUR feature set. The DUR model is then trained.

During training, all possible features per word (for PBS, PRM) or per phone (for DUR) were made available to the training algorithms. It was found in exploratory experiments that pre-selection of features did not significantly improve performance, and by using all available features the methodology becomes easier to port from language to language.

3. DATABASES

As part of this research, a suite of 6 databases (one per language) was designed, recorded and processed. Each database contains more than 2 hours (about 1400 sentences) of recorded speech, together with the following information (obtained fully- or semi-automatically): orthography, phonetic transcriptions, phonetic segmentation, prosodic labels (PRM and PBS) as well as linguistic features such as Part-of-Speech. These databases are described in more detail in [5].

For training and testing, the available data for each database was partitioned as shown in Figure 2. In contrast to A/B, C/D contains manually verified phonetic segmentation. Because A/B are used for word-level modelling and C/D for phone-level modelling, A/B are proportionally larger. Each test set contained at least 1000 data items (words in B, phones in D).

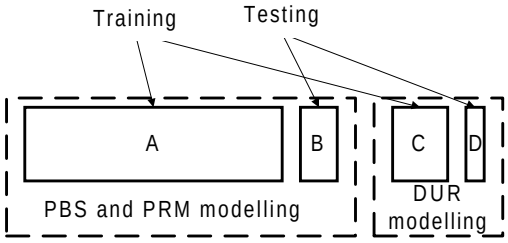

Figure 2 Data partitions used for training/testing.

4. RESULTS

4.1 Numerical Evaluation

For numerical evaluation, the predictions of the prosody models for PBS, PRM and DUR for the 'unseen' test data sets (Set B for PBS and PRM, Set D for DUR) were compared with manual labels (for PBS and PRM) and manual phone segmentation (for DUR). The results for all 6 languages are shown in Tables 2, 3 and 4.

Table 2 shows the results for PBS prediction. Since PBS has a heavily positively-skewed distribution (since most words have a PBS of 0), the table also shows the performance of a baseline predictor (which always predicts PBS=0). This baseline predictor is for some languages (French and Italian) *better* than the MLP and RT models. Of course the baseline predictor is not a viable model for a real TTS system, but it illustrates the fact that care must be applied when considering performance results on such skewed distributions. If a margin of error of ±1 is allowed, then both MLP and RT models predict more than 90% of PBS values correctly. The difference in performance between the two models is slight.

	EXACT			**EXACT ± 1**		
	'PBS=0'	**MLP**	**RT**	**'PBS=0'**	**MLP**	**RT**
Dutch	70.1	72.3	72.7	85.6	94.9	94.7
English	60.5	65.2	65.6	85.0	95.5	94.7
French	75.2	74.2	71.4	81.4	91.0	91.3
German	70.0	74.8	72.7	85.3	96.3	96.3
Italian	79.6	78.2	79.1	87.2	97.0	97.4
Spanish	86.9	88.7	89.7	93.2	97.3	97.3

Table 2 PBS-predicting performance of baseline, MLP and RT predictors on Test Set B: cell entries show the % of words for which the prediction is exactly the same as the observed value ('EXACT') or within ±1 of the observed value ('EXACT±1').

The PRM prediction results in Table 3 show significant variation between languages, but again fairly small differences between MLP and RT models. These results are based on cascade models of either two MLPs or two RTs.

	MLP	RT
Dutch	72.1	72.8
English	69.9	72.9
French	76.9	81.4
German	74.5	74.8
Italian	80.0	80.3
Spanish	90.8	92.2

Table 3 PRM-predicting performance of MLP and RT predictors on Test Set B: % of words for which model prediction is within ±1 of the observed value.

Table 4 shows the DUR prediction results. These results are based on cascade models of either three MLPs or three RTs. The correlation between observed phone durations and model durations is generally high, although the MLP models are consistently better than the RT models.

	MLP	RT
Dutch	0.80	0.79
English	0.78	0.75
French	0.73	0.69
German	0.78	0.75
Italian	0.84	0.83
Spanish	0.75	0.72

Table 4 DUR-predicting performance of MLP and RT predictors on Test Set D: Pearson correlation coefficient between the phone duration predictions of the model and the observed values.

4.2 Perceptual Evaluation

In order to try to establish whether the automatically-determined prosody models represent an improvement over existing hand-crafted prosody models, the best performing models for two languages, Dutch and English, were used to predict the prosody of 18 sentences drawn from Set D. The same sentences were also synthesized by the L&H TTS-3000 Synthesizer, and a forced-preference listening test, with 12 native-speaking subjects per language, was used to obtain preference information about the stimuli. For each language, four versions of each stimulus sentence were synthesized as shown in Table 5.

Label	Preparation of training data	Source of prosody (PBS, PRM, DUR)
OBS	N/a	Measured from recording
AUT	Automatic	Cascade Model
MAN	Manual	Cascade Model
TTS	N/a	L&H TTS-3000 Model

Table 5 Origins of the four test stimuli.

Although the MLP and RT models predict the variables PBS (0,..,3) and PRM (0,..,9), the synthesizer uses the simpler binary variables PAUSE and SAC (Sentence Accent). In order to achieve compatibility with the TTS system, simple threshold rules were used to convert the PBS and PRM values for the AUT, MAN and OBS models to PAUSE and SAC.

For each of the 18 sentences, 6 stimulus-pairs were formed covering every possible combination of the 4 prosody sources (OBS, AUT, MAN, TTS). Thus each listener heard 18*6=108 stimulus-pairs. The order of presentation of each stimulus was balanced across sentences (for each sentence, half the subjects heard the pair OBS-TTS while the other half heard TTS-OBS), and across the stimuli used for each subject (each subject heard OBS first in 27 cases, and heard OBS second in 27 cases (27=108/4)). The order of presentation of the actual pairs was randomized, ensuring that stimulus-pairs originating from the same sentence were separated from each other. For each stimulus-pair the subject was first shown the sentence text, then heard stimulus 1, then heard stimulus 2, and then chose a preference within 10s. If no preference was made within 10s, the stimulus-pair was marked as a 'timeout'.

Table 6 shows the percentages of preferences for each pairing for Dutch. For example, for stimulus-pairs containing the TTS and OBS prosody, 36% of the time the TTS model was preferred, and 63% the OBS model was preferred (the remaining 1% is accounted for by timeouts). The OBS model is significantly preferred to the MAN and TTS models (at the 0.01 level), but the difference between OBS and AUT is not significant. The AUT prosody is preferred to the TTS prosody. These results are consistent with the following preference order:

OBS > AUT > MAN > TTS,

with the proviso that the differences between pairs OBS-AUT, AUT-MAN and MAN-TTS are not significant.

		Preferred Choice			
		OBS	**AUT**	**MAN**	**TTS**
Not preferred	**OBS**	-	46	40*	36*
	AUT	54	-	48	39*
	MAN	60*	52	-	44
	TTS	63*	61*	56	-

*Table 6 Percentages of preferences for one prosody source over another for Dutch. *=significant at 0.01 level.*

Table 7 shows the equivalent results for English. In this case the OBS prosody is clearly preferred to the three prosody models. However, no significant difference is found *between* the three prosody models. Thus the following preference order is postulated:

OBS > AUT, MAN, TTS.

		Preferred Choice			
		OBS	**AUT**	**MAN**	**TTS**
Not preferred	**OBS**	-	37*	35*	31*
	AUT	63*	-	50	50
	MAN	65*	50	-	49
	TTS	69*	50	51	-

*Table 7 Percentages of preferences for one prosody source over another for English. *=significant at 0.01 level.*

Figure 3 shows a summary of the preference tests for both languages. For example, for stimulus-pairs containing the OBS prosody, the OBS stimulus was preferred 59% of the time for Dutch, and 66% of the time for English. The figure confirms, unsurprisingly, that the OBS prosody is the best, so there is still much room for improvement in the development of prosody models. However, both the AUT and MAN models offer improvements over the existing prosody generated within the TTS system.

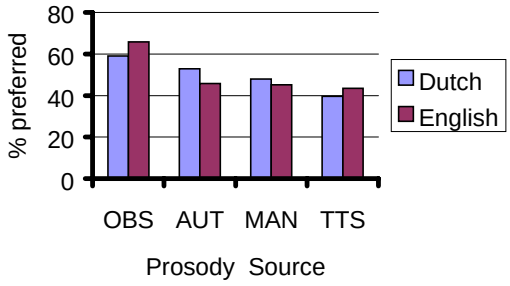

Figure 3 Summary of preference results.

5. CONCLUSIONS

The numerical evaluation results suggest that the cascade architecture of prosody used in this investigation is capable of producing relatively good predictions of three important prosodic variables. Although the integration of these into our existing TTS framework sacrifices some of the detail in these variables, these automatically generated models are generally preferred to the existing TTS prosody. It is found that the models trained on automatically labelled data are actually preferred to those trained on manual labels. This, together with the fact that the automatic labelling strategy is known to be effective for 6 languages [7], suggests that great efficiency improvements can be made in the development of prosody models for new languages.

6. ACKNOWLEDGEMENTS

This research was performed with support of the Flemish Institute for the Promotion of Scientific and Technological Research in Industry (contract IWT/AUT/950056). The authors would like to acknowledge the contributions made to this research by Lieve Macken, Ellen Stuer and Cynthia Grover.

7. REFERENCES

[1] Riedi, M. (1995), A neural-network-based model of segmental duration for speech synthesis. *Proceedings of Eurospeech'95*, Madrid, Spain, pp.599-602.

[2] Riley, M. (1992), Tree-based modellling of segmental durations. In: G. Bailly, C. Benoit and T.R. Sawallis (eds.), *Talking Machines: Theories Models and Designs*, Elsevier, Amsterdam, pp.265-273.

[3] van Santen, J., C. Shih, B. Möbius, E. Tzoukermann and M. Tanenblatt (1997), Multi-lingual duration modeling, *Proceedings of Eurospeech'97*, Vol 5, pp.2651-2654.

[4] Widera C., T. Portele and M.Wolters (1997), Prediction of word prominence, *Proceedings of Eurospeech'97*, Rhodes, Greece, pp.999-1002.

[5] Grover, C., J.Fackrell, H.Vereecken, J.-P.Martens and B. Van Coile (1998), Designing prosodic databases for automatic modelling in 6 languages, *Proceedings of ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp.93-98.

[6] Vereecken, H., A. Vorstermans, J.-P. Martens and B. Van Coile (1997), Improving the phonetic annotation by means of prosodic phrasing, *Proceedings of Eurospeech'97*, Rhodes, Greece, Vol 1, pp.179-182.

[7] Vereecken, H., J.-P. Martens, C. Grover, J.Fackrell and B. Van Coile (1998), Automatic Prosodic Labeling of 6 Languages, *Proceedings of ICSLP'98*, Sydney, Australia, Vol 4, pp.1399-1402.

[8] Grover, C., B. Heuft and B. Van Coile (1997), The Reliability of Labelling Word Prominence and Prosodic Boundary Strength, *Proceedings of ESCA Workshop on Intonation : Theory, Models and Applications*, Athens, Greece, pp.165-168.

[9] Breiman, L., J. Friedman, R. Olshen and C. Stone (1984), *Classification and Regression Trees*, Wadsworth International, Belmont, California, USA.

[10] Campbell, N. (1989), Syllable-level Duration Determination, *Proceedings of Eurospeech'89*, Paris, France, pp.698-701.