Combining Models of Prosodic Phrasing and Pausing

Tina Burrows, Peter Jackson, Katherine Knill, Dmitry Sityaev

Speech Technology Group, Cambridge Research Laboratory,
Toshiba Research Europe Ltd, 1, Guildhall St, Cambridge, UK
{tina.burrows, peter.jackson, kate.knill, dmitry.sityaev}@crl.toshiba.co.uk

10.21437/Interspeech.2005-557

Abstract

This paper describes two approaches to assigning prosodic phrase structure and pauses to text and investigates the impact of errors in the assignments for different granularities of prosodic phrase structure. One approach uses a cascaded combination of models trained separately for prediction of prosodic phrase structure and pauses and the other uses a model trained for the joint prediction task directly. Objective measurements show similar performance for both approaches while perceptual evaluations show a slight preference for an optimised cascaded combination of prosodic phrase structure and pause models using a single-level encoding of prosodic phrase structure.

1. Introduction

In Text-to-Speech (TTS) systems, the assignment of prosodic phrase structure and pauses to text is important to aid intelligibility and naturalness and to convey the underlying meaning of the text. Punctuation may help to indicate a writer's intended meaning or desired phrasing and is thus an important feature for both prosodic phrase boundary and pause prediction models. In speech, the position and strength of prosodic phrase boundaries (or 'breaks') are conveyed by a number of acoustic features such as pauses, lengthening of phrase-final syllables and changes in pitch. Stronger prosodic phrase boundaries (also referred to as primary or intonational phrase boundaries or breaks) are often marked by pauses and thus the strength of prosodic break at a word juncture is an important feature for pause prediction models.

Previous research on the assignment of prosodic phrase structure to text has used machine learning techniques such as decision trees [1, 2] and memory-based learning [3] trained on features derived from the text and its syntactic structure. While these methods have been fairly successful for identifying intonational phrase boundaries or just the location of phrase boundaries, assigning prosodic phrase boundaries of differing strengths has proved a more difficult task [4].

In this paper, decision trees are trained to assign prosodic breaks and pauses to text using features derived from the text and its syntactic analysis. The performance of models trained on a single-level encoding of prosodic phrase structure (indicating the presence but not strength of a phrase boundary) is compared to that of models trained on a multi-level encoding (which distinguishes between two strengths of phrase boundary). Motivated by the fact that prosodic break prediction and pause prediction are closely related and can be successfully modelled using similar features, models for the joint prediction of prosodic phrase boundaries and pauses are also trained. Their performance is compared with optimised combinations of prosodic break and pause models for both granularities of prosodic phrase structure. The impact of prosodic break prediction errors on the performance of models is also examined.

2. Experimental Setup

2.1. Data

Experiments were carried out on a 34525 word corpus of US English (one female speaker) which was collected by Toshiba Research Europe Ltd and annotated with ToBI tone and break tier annotation [5], syntactic mark-up (part-of-speech tagging and dependency relations), syllabification of words and position of pauses. The corpus consists of 2734 sentences of read text, covering a variety of sentence types (declaratives, questions, imperatives) and domains. In ToBI, there are 5 basic levels in the break tier annotation (0-4), corresponding to decreasing coupling between adjacent words and an increase in prosodic phrase boundary strength. For these experiments, levels 0, 1 and 2 were grouped together as 'non-breaks' (no prosodic phrase boundary) and levels 3 and 4 were considered as 'breaks' (phrase boundaries). The corpus contains 13614 breaks (4243 at level 3, 9371 at level 4) and 6779 pauses, with 40.3% of word junctures marked by either a pause or a break or both, and 0.7% of word junctures marked only by a pause. 0.7% of word junctures are marked by a level 3 break and a pause, compared to 18.2% of word junctures which are marked by a level 4 break and a pause (7.9% at the end of sentences). For training and testing, 10% of the data was withheld as an evaluation set and the remaining 90% was split into 10 cross-validation sets, with 90% for training and 10% for development. Sentence selection was randomised within each domain and sentence type so that the training, development and evaluation sets had a similar composition to the whole corpus.

2.2. Decision tree models

Models were trained for prosodic break prediction, for pause prediction (position only) and for the joint prediction of prosodic break and pause assignment. The models predict the relevant tag (or combined prosodic/pause tag) for each word juncture in an utterance. Pause tags represent a 'pause' or 'no pause' at a word juncture. Prosodic tags represent a 'break (one or more strengths of prosodic phrase boundary) or 'no break' at a word juncture. For 'breaks', two granularities of prosodic phrase structure were used; a single-level encoding, in which ToBI break levels 3 and 4 were merged together as a single 'break' tag (identifying the presence of a phrase boundary but not its strength) and a multi-level encoding in which the distinction between ToBI levels 3 and 4 was maintained by two separate prosodic tags.

Decision tree models were trained on the 10 cross-validation data sets for both granularities of prosodic phrase structure using C4.5 with subsetting [6]. The same feature set

was used for all tasks and included punctuation, part-of-speech and grammatical role (over a 5-word window centered on the current word), and the prosodic tags for the previous 5 word junctures. For pause models, the prosodic tag for the juncture following the current word was also included in the feature set.

2.3. Performance assessment

The feature set contains a window of prosodic tags for previous word junctures. For prosodic break models and models of the joint prediction task, two modes of operation are considered for performance assessment: *prediction,* in which a model's predictions for previous prosodic breaks provide the features for previous prosodic breaks, and *training,* in which the values of these features are taken from the corpus annotation. The test results reported by C4.5 training on a specified test set are generated in this mode and represent the best performance achievable on that data with an ideal input (no feedback errors).

For the feature set used in this work, there are no features derived from previous pause predictions, so for pause models in stand-alone operation, *training* and *prediction* modes are identical, as all features are derived from the corpus annotation in both cases. In practice in a TTS framework, pause predictions would be made by a pause model operating in a cascade combination with a prosodic break model, where prosodic break features are derived from the output of the prosodic break model with which the pause model is combined. Thus the pause prediction performance for the combination is affected by break prediction errors.

Quantitative performance is measured by precision (P), recall (R) and f-score (F) using a β-value of 1 [7]. For scoring the joint prediction task, a word juncture with no prosodic break or pause is the non-event and all other tag combinations represent a prosodic event to be scored.

Performance is compared with a baseline model which assigns prosodic phrase breaks (at level 4 for the multi-level baseline) and pauses to punctuation in the set (){}[];:,.!? The statistical significance of differences in performance between models is tested at a 95% confidence level using an exact two-sided matched-pair randomisation test [8], which allows for the lack of independence between each matched-pair due to overlapping cross-validation training data.

3. Results

3.1. Prosodic phrase break models

The average performance from 10-fold cross-validation models on various data sets is shown in Table 1. The high precision of the baseline model reflects the usefulness of punctuation as a feature for the prediction of prosodic breaks, mainly for ToBI level 4. All trained models have better recall and f-score than the baseline model. The lower performance of models using a multi-level encoding of prosodic phrase structure reflects the fact that it is a harder task to predict different strengths of prosodic phrase boundary. Average f-score from 10-fold cross-validation models on the evaluation data for prediction of ToBI level 4 is 80.5%, but only 29.1% for level 3, which reduces the overall f-score to 65.6%.

The effect of feedback of previous prosodic break predictions via the features is to reduce performance slightly, shown by the mean differences between prediction and training performance (prediction minus training) on the 10-fold cross-validation development sets (P, R and F respectively):

- Single-level: -0.3%*, -0.4%*, -0.3%*
- Multi-level: -0.2%, -0.1%, -0.1%

Although some are statistically significant (*), the differences are small.

Model, Mode,Data	Single-level			Multi-level		
	P	R	F	P	R	F
pro,trn,dev	83.4	79.6	81.5	67.2	61.7	64.3
pro,prd,dev	83.0	79.3	81.1	67.0	61.6	64.2
pro,prd,evl	81.8	79.7	80.8	67.7	63.5	65.6
pro,pnc,evl	100	33.7	50.4	98.5	33.2	49.6
pau,trn,dev	82.3	73.2	77.5	84.3	76.3	80.1
pau,prd,evl	81.2	77.7	79.4	85.0	80.6	82.7
pau,pnc,evl	95.6	68.3	79.7	-	-	-

Table 1: Average performance (%) from 10-fold cross-validation models for prosodic break and pause prediction on specified data sets. Model: 'pro'=prosodic break models, 'pau'=pause models. Mode: 'pnc'= baseline (at punctuation), 'trn'=test results reported by C4.5 training, 'prd'=results in prediction. Data: 'evl'=evaluation set, 'dev'= 10-fold cross-validation development sets.

3.2. Pause prediction models

Table 1 shows that pause models trained using a multi-level encoding of prosodic phrase structure perform better in training and stand-alone prediction than those trained using a single-level encoding. Mean differences in stand-alone prediction performance (multi-level minus single-level) on the evaluation set are 3.8%, 2.9% and 3.3% for P, R and F respectively. All are statistically significant and large enough to suggest that a multi-level encoding may be of practical importance for cascade combinations.

The baseline model performs fairly well and has a high precision because most punctuation (and all sentence-final punctuation) is marked by a pause. The models trained using a single-level encoding of prosodic phrase structure have a better balance between precision and recall than the baseline model, but the resulting f-score is very similar for prediction of the evaluation data.

3.3. Cascading prosodic phrase break and pause prediction models

To assess the impact of errors in prosodic break predictions on the performance of pause models when they are cascaded with prosodic break models in a TTS framework, each 10-fold cross-validation pause model was combined with the prosodic break model trained on the same cross-validation training set and the model cascade used to predict the evaluation data. The average performance for the 10-fold cross-validation cascades is shown in Table 2. Mean differences between the performance of the pause models in a cascade (in which prosodic break features are derived from the predictions of the prosodic break model) and the stand-alone performance (in which prosodic break features are derived from the corpus annotation) are:

Model	Prosodic Break			Pause			Combined tags		
	P	R	F	P	R	F	P	R	F
casc_s	81.8	79.7	80.8	80.2	76.8	78.5	67.5	64.5	66.0
joint_s	82.9	78.7	80.8	79.5	77.7	78.6	68.9	64.3	66.5
casc_m	67.7	63.5	65.6	79.9	78.0	78.9	58.5	53.8	56.1
joint_m	67.8	62.7	65.2	78.3	77.2	77.7	58.0	52.7	55.2
opt_casc_s	83.7	80.6	82.2	85.1	78.1	81.4	71.2	67.2	69.2
best_joint_s	84.2	80.3	82.2	79.9	78.9	79.4	71.3	66.9	69.0
opt_casc_m	69.3	64.4	66.8	80.7	79.2	79.9	61.1	55.7	58.3
best_joint_m	67.0	63.4	65.1	78.7	77.8	78.3	59.1	54.8	56.9

Table 2: Average performance (%) from 10-fold cross-validation models for prediction of the evaluation set by cascades of prosodic break and pause models trained on the same cross-validation data sets ('casc') and by models of the joint prediction task ('joint'). '_s' and '_m' indicate single-level and multi-level encoding of prosodic phrase structure respectively. Also shown are the best performances on combined tags from optimised cascades ('opt_casc') and joint models ('best_joint'), where 'best' indicates the highest f-score for prediction of the evaluation data.

- Single-level: -1.0%*, -1.0%*, -1.0%*
- Multi-level: -5.1%*, -2.6%*, -3.8%*

All differences are statistically significant (*) and show that pause prediction performance is reduced by prosodic break prediction errors. The loss in performance is greater for a multi-level encoding of prosodic phrase structure than for a single-level encoding and is large enough to be of practical importance in a cascaded combination. For pause prediction on the evaluation data, the mean differences between performance of cascades using a multi-level encoding and cascades using a single-level encoding (multi-level minus single-level) are -0.4%, 1.2%*, and 0.4% for P, R and F respectively. These differences are no longer statistically significant (except *), showing that much of the benefit of using a multi-level encoding of prosodic phrase structure for pause prediction has been cancelled out by the higher error-rate of prosodic break models predicting a multi-level encoding.

3.4. Joint models

3.4.1. Performance on combined tags

Joint models were trained to predict the combined tags directly to compare their performance to cascaded models. Table 2 shows that average cross-validation performances from joint models are very similar to those from cascaded prosodic break and pause models trained on the same data sets. The mean differences in performance between joint models and cascaded models (joint minus cascade) for prediction of combined tags on the evaluation data (P, R and F respectively) are:

- Single-level: 1.4%*, -0.2%, 0.5%
- Multi-level: -0.5%, -1.1%*, -0.8%*

The statistically significant differences (*) do not identify one approach as significantly better or worse than another overall, since joint models using a single-level encoding have better precision, but joint models using a multi-level encoding have worse recall and f-score on average than cascaded models. Joint models also showed a small loss in performance on combined tags due to feedback of errors in predictions of previous breaks via the model features shown by negative mean differences between prediction and training performance for combined tags on development data sets. Although statistically significant for a single-level encoding, differences were small (less than or equal to 0.3%).

Table 2 also shows the best combined tag performances obtained from joint models and from cascades by optimising the models used in the combinations. For each level of encoding, the cascades were optimised by first selecting the prosodic break model (from all the cross-validation models trained) with highest f-score for break predictions on the evaluation set and then selecting the pause model to maximise the f-score for the combined tags from the cascaded models. Nearly all performance measures for the optimised cascades are better than those for the best joint models and all f-scores are better (except for single-level break prediction which is equal).

3.4.2. Performance on prosodic break tags

Prosodic break prediction scores were calculated for the joint models from the predicted prosodic/pause tag combination. The average break prediction performances from joint models are very similar to those from prosodic break models. Mean differences in performance measures between joint models and prosodic break models trained on the same data sets (joint minus prosodic break) for prosodic break prediction on the evaluation set (P, R and F respectively) are:

- Single-level: 1.1%*, -1.0%*, -0.0%
- Multi-level: 0.1%, -0.8%*, -0.4%

The differences of statistical significance (*) may be large enough to be of practical importance. The f-score is not significantly different for either approach for either level of encoding. For a single-level encoding, the optimised cascade and best joint model have the same f-score.

3.4.3. Performance on pause tags

Pause prediction scores were calculated for the joint models from the predicted prosodic/pause tag combination. Joint models using a single-level encoding have similar average pause prediction performances to those of cascaded models, but models using a multi-level encoding have significantly worse f-score. This is shown by the mean differences between their pause prediction performances (joint minus cascade) on the evaluation data (P, R and F respectively):

- Single-level: -0.8%, 0.9%*, 0.1 %
- Multi-level: -1.6%, -0.8%, -1.2%*

These differences are statistically significant (*) for single-level recall (better) and multi-level f-score (worse).

Comparing the pause prediction performance from joint models trained using single-level and multi-level encodings of prosodic phrase structure, the mean differences on the evaluation data (multi-level minus single-level) are -1.20%*, -0.42% and -0.81%* (P, R and F respectively). The statistically significant differences (*) for precision and f-score show worse pause prediction performance for joint models trained using a multi-level encoding of prosodic phrase structure.

4. Perceptual Evaluations

The optimised model cascades and best joint models which maximised combined tag f-score on the evaluation data (Table 2) were used for perceptual evaluations. A set of 50 sentences were chosen from the evaluation data such that for each sentence chosen, at least one word juncture within the sentence had combined tag predictions that differed for all of the four models compared. The selected sentences covered a range of sentence lengths (6-24 words) and prosodic phrase structures. Sentences were synthesised using Toshiba's US English TTS system, using the corpus annotation for the features derived from syntactic analysis. Predicted pauses were assigned fixed durations, the durations varying depending on whether the word juncture was also marked by punctuation or a predicted prosodic break. Evaluation of all pairwise model combinations was performed by six subjects whose dominant language was English, for a total of 900 comparisons per model. The order of presentation of sentences, and order of models within each pair, was randomised in each evaluation. The overall ranking based on the preference counts was:

opt_casc_s>best_joint_m >opt_casc_m > best_joint_s

The significance of differences between models was tested using a paired Wilcoxon signed rank test and showed that an optimised model cascade using a single-level encoding of prosodic phrase structure was significantly preferred overall but that there was no significant preference between all other models.

5. Conclusions and Further Work

Two approaches to assigning prosodic phrase structure and pauses to text in a TTS system have been presented in this paper and the impact of the granularity of prosodic phrase structure representation on performance was considered. Objective assessments showed that the average performance from cascades of models trained separately for prosodic break prediction and pause prediction was very similar to the average performance of models trained for the joint prediction task directly. However, optimised cascade combinations could outperform the best joint models scored on the combined tags. Perceptual evaluations showed a slight preference for an optimised cascade combination trained with a single-level encoding of prosodic phrase structure.

Pause model performance was shown to be improved by using features with more granularity in prosodic phrase structure representation. However, for a cascade combination of pause and prosodic break models, for example as in a TTS system, there was no longer a clear benefit for pause prediction performance from a multi-level encoding of prosodic phrase structure. This is because predicting a multi-level encoding of prosodic phrase structure is a more difficult task for the prosodic break models and errors in prosodic break predictions are propagated to the pause model via the feature set. Predicting level 3 breaks is particularly difficult, since they occur much less frequently in the data (only 12.3% of word junctures are marked by a level 3 break compared to 27.1% by level 4). For models of the joint prediction task, pause prediction performance was not improved by using a multi-level encoding of prosodic phrase structure.

For a TTS system, joint models provide the advantage of a faster one-step prediction of prosodic and pause tags, but have the disadvantage that the same feature set is used for both prosodic break and pause prediction and features cannot be tuned for each task individually. Further, pause predictions cannot utilise future context of prosodic phrase boundary information. Preliminary experiments on pause prediction using additional features for prosodic breaks for 5 word junctures following the current word juncture did not show a significant difference in performance however, suggesting such features may not be that important.

Further work to investigate the impact of feedback of pause predictions via the feature set would be interesting, for example by including features for a window of previous pause predictions or a measure of the distance to the previous pause. Further perceptual experiments would also be useful to assess how the trade-off between precision and recall affects perceptual evaluations and to determine the best model selection criteria, for example whether to maximise performance for the joint prediction task or for each task sequentially.

6. References

[1] Wang, M. W., Hirschberg, J., "Automatic classification of intonational phrase boundaries", *Computer Speech and Language*, 6:175-196, 1992.

[2] Koehn, P., Abney, S., Hirschberg, J., Collins, M., "Improving intonational phrasing with syntactic information", in *Proceedings ICASSP-2000*, Istanbul, 2000.

[3] Busser, B., Daelemans, W., van den Bosch, A., "Predicting phrase breaks with memory-based learning", in *Proceedings 4th ICSA Tutorial and Research Workshop on Speech Synthesis*, pp. 29-34, Scotland, 2001.

[4] Ingulfsen, T., "Influence of syntax on prosodic boundary prediction", *Tech. Rep. UCAM-CL-TR-610*, University of Cambridge Computer Laboratory, 2004.

[5] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., "ToBI: A Standard for Labeling English Prosody", in *Proceedings of the International Conference on Spoken Language Systems*, 2: 867-870, Banff, Canada, 1992.

[6] Quinlan, J. R., *C4.5: Programming for Machine Learning,* Morgan Kaufman, San Mateo, CA, 1993.

[7] Van Rijsbergen, C. J., *Information Retrieval,* Buttersworth, London, 1979.

[8] Edgington, E. S., *Randomization Tests*, 3rd Ed., Marcel Dekker, New York, 1995.