Adaptation of Prosodic Phrasing Models

Peter Bell [*]   *Tina Burrows* [†]   *Paul Taylor* [‡]

peter.bell@ed.ac.uk   tina.burrows@crl.toshiba.co.uk   pat40@cam.ac.uk

Abstract

There is considerable variation in the prosodic phrasing of speech between different speakers and speech styles. Due to the time and cost of obtaining large quantities of data to train a model for every variation, it is desirable to develop models that can be adapted to new conditions with a limited amount of training data. We describe a technique for adapting HMM-based phrase boundary prediction models which alters a statistical distribution of prosodic phrase lengths. The adapted models show improved prediction performance across different speakers and types of spoken material.

1. Introduction

A key element of prosody is the division of an utterance into prosodic phrases, indicated in speech by pauses, changes in amplitude and pitch, and the lengthening of the final syllable within a phrase. This helps to disambiguate the meaning of an utterance. The prediction of phrase boundaries is an important task for text-to-speech synthesis applications, not only to convey the correct meaning, but also to increase the naturalness of the synthesised speech.

Despite a correspondence with syntactic structure, there is considerable variation in prosodic phrasing between speakers and types of speech. For example, in fast speech there will be fewer pauses: speakers tend to concatenate the phrases that are least significant for conveying meaning. Conversely, in very slow speech, phrasing may be shorter: there may be pauses between every significant word. In addition, phrasing is likely to vary with the domain of the subject material. For example, in the reading of text with very long sentences, the reader could make greater use of phrasing to enable the text to be more readily understood. For non-rehearsed news-reading, the presenter may opt for a style of regular phrase lengths with a lower dependence on the text itself, allowing the speech to sound interesting but avoiding the challenge of deep-parsing the text on the fly.

This variability presents challenges for the design of phrase break prediction algorithms if synthesised speech is to sound appropriate for its intended domain. It is desirable, given the difficulty and cost of obtaining large quantities of training data for every desired speaking style and text type, to develop phrasing models that can be easily adapted to new conditions with a reduced amount of training data, but minimal reduction in predictive power.

2. Prediction Model

2.1. Description

Our model is based on the HMM-based approach of Taylor and Black [1], extended by Schmid and Atterer [2]. We assume that the variability in phrasing is due to the underlying distribution of phrase lengths in a given speech domain and that this distribution can be adapted to the new domain.

Given a sequence of words, we aim to predict a sequence of breaks and non-breaks corresponding to junctures between words. Following [2], we do not distinguish between intermediate and full intonational phrase boundaries, and attempt to find the juncture sequence $J = (j_1, \ldots, j_n)$ giving the highest likelihood of the words, each juncture j_i taking the value B (break) or N (non-break). We set the "context" C_i to be the set of features, extracted from surrounding words, considered relevant to the prediction of the value of j_i: in common with other approaches such as [3], [4], these are the parts-of-speech of the two words immediately prior to the juncture and one following. We therefore seek to maximise

$$\hat{J} = \arg\max_J p(J|C) \quad (1)$$

$$= \arg\max_J \frac{p(C|J)P(J)}{p(C)} \quad (2)$$

We ignore the denominator $p(C)$ since this is invariant to change in J.

Defining d_i to be some measure of the "time" elapsed since the most recent break prior to the i^{th} juncture, j_i – the length of the phrase up to that point, we make the following simplifying assumptions:

- The Conditional Independence assumption,

$$p(C_i|J) = p(C_i|j_i) \quad (3)$$

- The existence of a *phrase-length* model:

$$p(j_i|J) = p(j_i|d_i) \quad (4)$$

  This states that the probability of a juncture, ignoring context, depends only on the time since the last break, rather than on the exact sequence of preceding junctures.

We find that the best performance is achieved when d_i is measured as the number of intervening syllables, as a proxy for word duration. Under these assumptions, (2) becomes

$$\hat{J} = \arg\max_J \prod_i p(C_i|j_i)p(j_i|d_i) \quad (5)$$

2.2. Viterbi Formulae

Equation 5 can be solved by making the Viterbi approximation: we assume that only the most likely sequence of junctures contributes to the likelihood of the observed features. We define the

[*] Centre for Speech Technology Research, University of Edinburgh, 2 Bachleuch Place, Edinburgh, UK

[†] Speech Technology Group, Cambridge Research Laboratory, Toshiba Research Europe Ltd, Cambridge, UK

[‡] Department of Engineering, University of Cambridge, Trumpington St, Cambridge, UK

path probability $\phi_d(i)$ to be the probability of the *most likely* partial juncture sequence $(j_0, j_1, \ldots, j_i)$, ending with a phrase at least d syllables long. $|w_i|$ is the length, in syllables, of the word immediately preceding the i^{th} juncture and $\phi_d(0)$ (given in [2]) is:

$$\phi_d(0) \quad = \quad \begin{cases} 1, & d = |w_1| \\ 0, & d \neq |w_1| \end{cases} \tag{6}$$

And $\phi_d(i)$ takes the values

$$\begin{array}{ll} 0 & d < |w_{i+1}| \\ \max\limits_{1 \leqslant k \leqslant D} \phi_d(i-1)p(B|k)p(C_i|B) & d = |w_{i+1}| \\ \phi_{d-|w_{i+1}|}(i-1)p(N|d-|w_{i+1}|)p(C_i|N) & d > |w_{i+1}| \end{array} \tag{7}$$

After the equations have been updated for every word in the text, we use a traceback procedure to find the optimal juncture sequence. To do this, a token-passing approach is used. Tokens store the position of the most recent break for the path considered, and are propagated forwards for $d > |w_{i+1}|$. For $d = |w_{i+1}|$ (signifying that juncture i is a break, B) the token corresponding to the best value of k, as used in the update equations, is recorded. The best sequence of breaks can then be found by working backwards using these recorded values.

2.3. Parameter Estimation

To estimate the probabilities $p(C_i|j_i)$ we use the C4.5 classifier [5] to create a decision tree from training data according to an entropy gain criterion. We identify the leaf L corresponding to features C_i. The tree stores the counts of training cases of B and N at L, allowing us to estimate

$$p(B|C_i) = \frac{\#B \ at \ L}{total \ cases \ at \ L} \tag{8}$$

and the context-independent probability $P(B)$ is estimated from total counts over the whole tree. We then use Bayes rule to obtain

$$p(C_i|B) = \frac{p(B|C_i)p(C_i)}{p(B)} \tag{9}$$

Note that $p(C_i)$ can be ignored since it does not vary with the sequence of junctures. $p(C_i|N)$ can be estimated similarly. The decision tree approach avoids data sparsity issues by grouping together similar contexts. In the event that there are insufficient training cases at L to obtain a reliable estimate it is possible to back off to L's parent node in the tree, obtaining juncture counts from this node instead.

To estimate $p(j_i|d_i)$ we obtain counts of breaks, $C_B(d)$, and non-breaks, $C_N(d)$, occurring at a distance d from the previous break, by processing the training data sequentially. Then, simply,

$$p(B|d) = \frac{C_B(d)}{C_B(d) + C_N(d)} \tag{10}$$

Syllable counts are obtained using a heuristic algorithm. Data sparsity problems are very small compared with an N-gram model (such as that proposed by [1]) where the number of parameters to estimate increases exponentially with D, rather than linearly as in this case. However, problems do occur at high d as the pool of training cases is much smaller: only phrases that are at least as long as d contribute a case to the calculation of $p(B|d)$.

3. Adaptation of Phrase-Length Model

We aim to adapt well-trained phrasing models to new domains, using a relatively small, fixed amount of labelled adaptation data. To do this, we find estimates of the $p(j_i|d_i)$ by adapting the values of $p(B|d)$, the probability of a break, given that the current phrase is already of at least length d.

Underpinning our approach is the assumption that each domain has its own intrinsic "average phrase length". This might vary with, for example, speaking rate – Yeon-jun Kim and Yung-hwan Oh [6] have observed that the higher the speaking rate, the more words there are in a prosodic phrase – or with the material spoken – a lengthy radio news story may have long, regular phrases, whilst simple directions from an in-car navigation system may have many more pauses to make the utterance as easy as possible to understand. To relate this notion to the instantaneous break probabilities used by the models, however, we need to consider the underlying phrase length distribution.

We define Y to be the discrete random variable representing the distribution of phrase lengths over a particular domain. The phrase length distribution function is

$$F(y) = p(Y \leqslant y)$$

and we set

$$\lambda_y = p(Y = y | Y \geqslant y)$$

This is the probability of a phrase break occurring at a juncture, given that the phrase is already y units long. Since every phrase must have a positive length, $p(Y \geqslant 1) = 1$ and $\lambda_0 = 0$.

The formula (10) for estimating $p(B|d)$ actually estimates the values given by

$$\begin{aligned} p(B|d = y) &= p(Y = y|Y \geqslant y, \text{ juncture at } y) & (11) \\ &= \frac{\lambda_y}{p(\text{juncture at } y|Y \geqslant y)} & (12) \end{aligned}$$

This calculation takes account of the fact that it is not possible for the phrase to end at any arbitrary y – only those which coincide with word endings (junctures). However, the denominator above can be simply approximated by the word-to-syllable ratio of the data, which was found to be close to 0.6 for all the datasets considered here. Therefore we ignore the distinction between $p(B|d = y)$ and λ_y.

It follows from this that

$$\begin{aligned} p(Y > y) &= p(Y > 1|Y \geqslant 1) \cdots p(Y > y|Y \geqslant y) & (13) \\ &= \prod_{k=0}^{y} (1 - \lambda_k) & (14) \\ \Rightarrow F(y) &= 1 - \prod_{k=0}^{y} (1 - \lambda_k) & (15) \end{aligned}$$

So we can derive an estimate for the distribution function $F(y)$ via estimates of λ_y, from the original estimates of $p(B|d)$, and vice versa. The formula can be inverted using

$$\lambda_y = \frac{1 - F(y)}{1 - F(y-1)} \tag{16}$$

although some smoothing is required as $F(y)$ increases towards 1, when this fraction is sensitive to small rounding errors.

Figure 1 shows graphs of the probability density functions for the phrase length r.v. Y, for three data from three different domains used in these experiments. No common type of probability distribution provides a sufficiently good fit to these

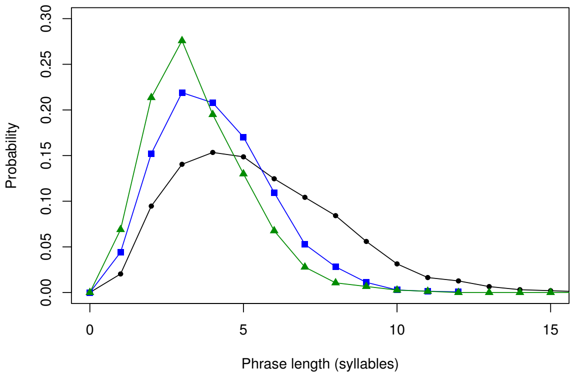

Figure 1: *Phrase length probability density functions for the F2B domain (dots), A-J domain (squares) and X-Z domain (triangles)*

functions (of those investigated, the closest was the Poisson distribution). However, to make an effective adaptation using a small amount of data from the new domain, we wish to reduce the distribution to a very small number of degrees of freedom. We therefore make the assumption that the underlying "shape" of the distribution is common to all domains, and that a linear transform to link any two can be found. This means that, given well-trained estimates for the phrase length distribution of training data, Y_{trn}, this can be transformed to one appropriate to a new domain using

$$Y_{new} = aY_{trn} + b \tag{17}$$

In fact, we know that $b = 0$ since all phrase length distributions are constrained so that $F(0) = 0$. Thus $\hat{a}$, the estimate of the true value for a can be simply determined using a method of moments, taking expectations of both sides:

$$\begin{aligned} \mathbb{E}(Y_{new}) &= a\mathbb{E}(Y_{trn}) \qquad (18) \\ \Rightarrow \hat{a} &= \frac{\hat{\mu}_{trn}}{\hat{\mu}_{adp}} \qquad (19) \end{aligned}$$

where $\hat{\mu}_{trn}$ and $\hat{\mu}_{adp}$ are estimated means derived from the observed phrase length distributions of the training data and adaptation data respectively.

Having estimated $\hat{a}$, we carry out the following:

- Calculate $\hat{F}_{new}(y) = \hat{F}_{trn}(y/\hat{a})$, using interpolation on the discrete function $\hat{F}_{trn}(y)$.
- Obtain, from this function, values of $\hat{\lambda}_y$ using (16) (smoothed for values of y above around 10), which can be used as estimates for $p(B|d = y)$ for the new domain.

A major advantage of this method is that it avoids data sparsity problems at higher d, where there are fewer training cases. If there is a limited amount of training data for a domain, $p(B|d)$ is nevertheless likely to be estimated accurately for low d, as there will still be a large number of training cases (one for every phrase that is at least length d). However, accuracy will rapidly diminish for higher d. The advantage of using the limited training data to adapt a well-trained phrase-length model, rather than to construct one independently, is that all training cases are used equally in the estimation of the average phrase length; the well-trained model, adjusted for this average, can be used to obtain a reliable estimate throughout the required range of d.

4. Evaluation

4.1. Experimental Setup

To source data from a range of domains, we use transcription data from the Boston University Radio News Corpus (BURNC) and the Toshiba US-English Female TTS Voice Corpus, kindly made available for this work by the Speech Technology Group, Toshiba Research Europe Limited. From the BURNC we use data from the F2B speaker which is fully annotated with the ToBI labelling scheme. This has been used for phrase break prediction by several others, such as [7] and [8]. The F2B set consists of 166 utterances and 13,075 junctures in total The Toshiba corpus consists of recordings from one female speaker, grouped into sets according to the type of material. Examples are shown in table 1. The sets we used, A-J, L-O and X-Z, contain 1592 utterances and 18,175 junctures in total. All are annotated with prosodic break tags. The F2B utterances tend to be much lengthier than those in any of the Toshiba categories, giving the greatest contrast.

Table 1: *Categories in the Toshiba Corpus*

Domain	Sets
Declaratives	A-J
Exclamations	K
Questions	Q
ICE Corpus sentences	L-O
Sentences for in-car navigation	X-Z

From the utterance transcriptions, sequences of POS tags (using the Penn Treebank POS tagset) were generated using the MXPOST tagger; these were used as inputs to the prediction model. 90% of each set was designated as training data. A portion of the remaining data (typically 1%-5%) was used as testing or adaptation data. The evaluation measures used were precision, the proportion of predicted breaks that are correct; recall, the proportion of breaks in the reference data that are correctly predicted; and F-score, the harmonic mean of precision and recall.

4.2. Results

The greatest inter-domain difference was found to be between the BURNC F2B set and the Toshiba X-Z set. We present results for adaptation experiments between these two domains, shown in table 3; other results were similar in trend, though less significant in size. A short description of the experiments is given in table 2.

Table 2: *Experiments*

Experiment	**Description**
Native	trained on full 90% X-Z set
Limited	trained on reduced 5% X-Z set
F2B	trained on full F2B data with no adaptation
Adapted	trained on full F2B data, adapted with 5% X-Z data

The performance of the HMM-based predictive models, trained on full set of native data for each domain, ranged be-

Table 3: *Results on the X-Z test set*

Phrase Model	Precision	Recall	F-Score
Native	79.5	83.5	81.5
Limited data	79.8	72.9	76.2
F2B	90.6	65.5	76.1
Adapted	85.0	74.3	79.3

Table 4: *Perplexity on the X-Z set*

Phrase Model	Perplexity
Native	1.708
Limited data	-
F2B	1.920
Adapted	1.715

tween F-scores of 80.1% and 83.3% and on the F2B set was 81.4%.

Comparing out-of-domain performance on prediction of data from the X-Z set, the model trained on full F2B data set gave an F-score of 76.1% (compared to 81.5% by the model trained on full X-Z training set). By applying the adaptation technique to the F2B model, performance was increased by 3.2% to 79.3%. This is better than the performance of a model trained on a limited amount of data from the native data set (76.2%).

The performance of the adapted models can be compared to that of the originals by calculating the perplexity on the test set. The perplexity figures for the phrase-length models shown in table 4 are inversely correlated with the performance, as would be expected. Graphs of $p(B|d)$ for the different models are shown in 2. Note that the curve for the model trained on a small amount of adaptation data is much less smooth, illustrating the unreliability of estimation, whilst the adapted curve is smooth and fits the data well at low d.

5. Summary

We have implemented a successful algorithm for data-driven prosodic phrase-break prediction, and presented a method for adapting it to a different type of speech or spoken material, using a limited amount of adaptation data from the new domain. The performance of the algorithm itself compares favourably with previous results. For example, the highest F-score achieved by [4] on the Boston corpus was 77.9%, compared to 82.1% here. The highest score achieved by Busser et al on data from the MARSEC corpus, was 78.3%. Our highest overall result was 83.3% on the Toshiba A-J set.

The adaptation technique investigated was found to be effective, when measured against performance both of models trained on foreign data and of models trained directly on the adaptation data. In general, the adaptation was found to result in a greater improvement when a smaller amount of adaptation data was used.

The data sets used here were not as heterogeneous as we would have liked: reductions in performance due to training on data from a different set were relatively small – no more than a fall in F-score of about 9% on the best models. A greater drop would probably have resulted in a wider spread of results using the different methods, giving a greater contrast between them. The only variations between the different domains as recorded were in the speaker and type of material – more interesting results might be generated using material recorded with a range of different speaking rates or different emotions. It would be particularly interesting to investigate the potential use of adapted phrase-length models for controlling tempo in synthesised speech, which could offer an improvement over methods that simply insert additional breaks after particular syntactic phrases when generating slower speech [9].

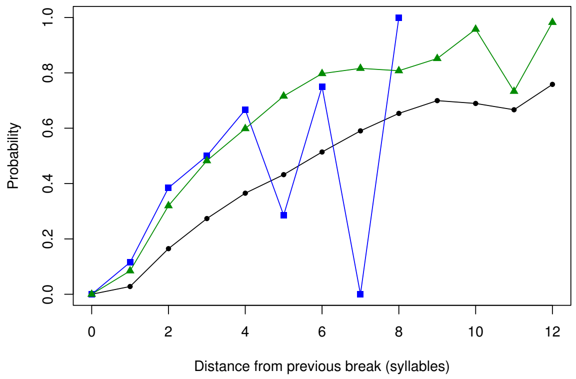

Figure 2: *$p(B|d)$ estimated from F2B data (dots), X-Z adaptation data (squares) and F2B data adapted using X-Z data (triangles)*

6. References

[1] P. Taylor and A. Black, "Assigning phrase breaks from part-of-speech sequences," *Computer Speech and Language*, vol. 12, pp. 99–117, 1998.

[2] H. Schmidt and M. Atterer, "New statistical methods for phrase break prediction," in *Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

[3] B. Busser, W. Daelemans, and A. van den Bosch, "Predicting phrase breaks with memory-based learning," in *Proceedings 4th ISCA tutorial and research workshop on speech synthesis*, 2001, pp. 29–34.

[4] T. Ingulfsen, "Influence of syntax on prosodic boundary prediction," Master's thesis, Cambridge University, 2004.

[5] J. R. Quinlan, *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1993.

[6] Y. J. Kim and Y. H. Oh, "Prediction of prosodic phrase boundaries considering variable speaking rate," in *Proceedings of the International Conference on Spoken Language Processing*, 1996, pp. 1505–1508.

[7] M. Fach, "A comparison between syntactic and prosodic phrasing," in *Proceedings of Eurospeech '99*, vol. 1, 1999, pp. 527–530.

[8] C. Fordyce and M. Ostendorf, "Prosody prediction for speech synthesis using transformational rule-based learning," in *Proceedings of International Conference on Spoken Language Processing*, 1998.

[9] J. Trouvain, "Tempo control in speech synthesis by prosodic phrasing," in *Proceedings "Konferenz zur Verarbeitung natrlicher Sprache"*, 2002, pp. 227–230.