Modeling Pause-Duration for Style-Specific Speech Synthesis

Alok Parlikar and Alan W Black

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA. USA
{aup, awb} @cs.cmu.edu

Abstract

A major contribution to speaking style comes from both the location of phrase breaks in an utterance, as well as the duration of these breaks. This paper is about modeling the duration of style specific breaks. We look at six styles of speech here. We present analysis that shows that these styles differ in the duration of pauses in natural speech. We have built CART models to predict the pause duration in these corpora and have integrated them into the Festival speech synthesis system. Our objective results show that if we have sufficient training data, we can build style specific models. Our subjective tests show that people can perceive the difference between different models and that they prefer style specific models over simple pause duration models.
Index Terms: Speech Synthesis, Style-specific Pause Duration, Phrasing

1. Introduction

Phrase breaks in natural speech are often realized with short silences. As the TOBI scheme [1] recommends, there can be different levels of phrase breaks. These different types of pauses have different duration values. For example, pauses at the ends of sentences are likely to be longer than those that occur within a sentence. Pauses are also style specific: they occur differently in audio books, say, than in broadcast news. In our previous work [2], we predicted placement of style-specific phrase breaks: where the breaks ought to be in an utterance. However, there is more to style than just the position of these breaks. It has been observed [3] that the length of individual pauses in speech are distributed differently for different individuals, as well as the type of situation in which speech is uttered. Duration of pauses also affects perception. For example, pause duration is a reliable means of discriminating between lexical ambiguity of words [4]. From a style-specific synthesis point of view, it is thus important to model not just where we insert prosodic breaks, but also to model the duration of these silences.

Although phrase break prediction has been widely explored in synthesis, generating these breaks with appropriate duration has not received much attention. Generally, all duration models treat the pause separately. Some segmental duration modeling techniques, such as [5] does not predict pauses at all. [6] divides an utterance into rhythmic groups and predicts the duration of each group. It computes the segmental duration of the group and then optionally inserts a pause of the remainder length. Following the Klatt model [7], the Festival speech synthesis system [8] as well as the Mary TTS system [9] assign fixed duration to breaks, based on the predicted TOBI level of the break. One of the reasons why pause specific duration models have not been thoroughly explored is that style corpora with appropriate annotations are not easy to construct. The BURNC corpus [10] is one such corpus, but it hasn't been used to build pause duration models.

This paper presents a data-driven approach to modeling duration of phrase breaks. Similar to [2, 11], we use forced-alignments between speech and transcriptions to detect where phrase breaks are in natural speech. For each break, we find out its duration and extract features that could be used to learn a regression predictor for the duration. Here we present our work on six data sets, which vary both in size as well as speech style. In this work, we only model breaks that occur within an utterance. We do not yet model the breaks at sentence boundaries in paragraph level utterances.

In the following sections, we shall first look at the different styles of speech used in this work. We shall then present analysis of natural break duration which shows that the speaking styles are indeed quite different. We shall then describe the set of features that we use to model the pause duration. We shall then look at the performance of style specific models and compare it to the performance of Festival's fixed-duration model as well as a generic duration model trained using our method on combined data of all the styles. Finally, we shall look at subjective evaluation of our models and discuss our results.

2. Corpora and Styles

The notion of speaking style is a nebulous one [12]. Different people have their own general speaking style. But the same person could adopt different styles when reading passages from different genres of text. Similarly, two different speakers may have very different style of reading the same genre of text. Speaking style can also vary depending on the task at hand. The style of read speech may also be different than that of spontaneous speech.

We looked at six speech corpora in this work. The *Europarl* corpus consists of prompts from the English side of the Europarl [13] parallel corpus between English and Portuguese. This data contains proceedings of the European Parliament. The speech was recorded by an Indian English speaker (AUP) in the style of "parliament proceedings". The *ARCTIC* corpus consists of the ARCTIC prompt set [14] recorded by speaker SLT (female, American speaker). The style of this speech is "short sentences". The *F2B* corpus is from the Boston University Radio News Corpus [10], in the style of "radio broadcast". The *Obama* corpus consists of public talks by the US President, Barack Obama. Audio and transcripts of two of his public addresses were used to build this voice: (i) Presidential Candidate speech (Mar 2008, Philadelphia) and (ii) Address at the Military Academy (Dec 2009). The *TATS* corpus is taken from an audio-book (The Adventures of Tom Sawyer, by Mark Twain) in the Librivox database. The book was recorded by a male professional volunteer. This is in the "audio-book" style. Finally, the *Emma* corpus [15] is taken from an audio-book (Emma, by Jane Austen) in the Librivox database. The book was recorded by a female volunteer. The style of this corpus is also, broadly, "audio-book" but is

different from the *TATS* corpus.

We extracted the pause duration from natural speech in our corpora. To do that, we force-aligned the speech and transcriptions using an EHMM tool [16] that allows for short silences to be inserted during the alignment. We used these alignments to find out the length of these inserted silences. We ignored all inserted pauses that were less than 80msec in length.

Table 1 shows the break duration profile of these corpora. Note that we do not include breaks at the end-of-utterance in our analysis here. This is because, for some databases, the end-of-utterance pause timing may no longer be in the database due to external pre-splitting, and/or recording being done as isolated sentences. We observe from this table, however, that the average break duration varies quite a bit across the styles.

Table 1: Break Profile of our corpora. Counts here do not include breaks at ends of utterances.

Corpus	Speech Size (minutes)	Num Breaks Per Minute	Average Break Length (msec)
Europarl	49	7.2	141
ARCTIC	56	3.8	130
F2B	55	10.6	273
Obama	61	11.8	414
TATS	406	7.6	249
Emma	1040	8.5	243

3. Analysis of Pause Duration

As argued in [17], it is better to use log-transformed duration, rather than values in the time domain, to analyze or model pauses. This is because corpus studies have shown that the log-distribution is closer to being a normal distribution than the original distribution. We looked at the log distribution of the duration of breaks extracted from our corpora.

Figure 1 plots the kernel density estimates of the log-distribution of breaks that we extracted. We observe that the distribution of breaks between different corpora is quite different. Also observe that even in the log domain, our break distribution is far from normality. We can perhaps consider that the break duration values come from a Gaussian mixture. The analysis in [17] described a trimodal distribution of pauses, and categorized them as brief (<200ms), medium (200-1000ms) and long (>1000 ms). In our analysis, we see that different styles have different modalities of duration.

4. Building a Duration Model

We started with the breaks extracted from each of our corpus. We dumped a set of features corresponding to those breaks, and then used a CART tree to model the prediction as regression on the training data. We held out 10% of the available data for testing. We split the remaining data into two: 90% for training and 10% for development. We trained the CART model using the wagon tool, with the stepwise option. This allowed us to select the most informative feature by evaluating it on the development set when building the tree.

Breaks only occur a few times in speech, and for all our corpora, we have only a small amount of data available to train regression from. We experimented with different stop values for CART training and eventually decided to use a low value of 5 items in every leaf node.

At synthesis time, Festival first predicts the positions of all breaks in an utterance and then builds the duration of segments

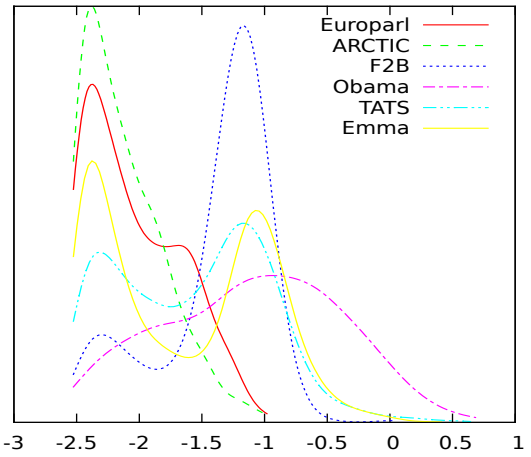


Figure 1: Kernel Density Plot of log-duration of breaks

from left to right. We used the standard Phrasing model in Festival[18] for the experiments here. When we encounter a segment that corresponds to a phrase break, we use the trained tree and predict the duration of that break.

The set of features that we used in our modeling is as follows: name of the two segments before and after the break, parts of speech of the two words before and after the break, punctuation character (if any) before the break, whether or not there is a quotation mark after the break, the number of content words in the previous phrase, and the number of stressed syllables in the previous phrase.

5. Evaluating a Duration Model

We can evaluate our pause-duration model in the same way that we typically evaluate other duration models in Festival. We use our held out test data as ground truth, and evaluate our prediction on that data using two measures: the RMSE of the predicted value, and the correlation between the actual duration and our predicted duration. Ideally we want to achieve a low RMSE number, and a high correlation number.

6. Experiments and Results

We have six corpora at hand. We extracted breaks and related features from each of the corpora. We then partitioned this data into 10 sets, with the intent of doing a 10-fold cross validation. In every set, we held out 10% of the data for testing. Instead of taking every tenth item into our test set, we preserved the sequence of breaks in training and testing data.

For each cross-validation fold and for each corpus, we have four different models. The baseline model is what Festival uses by default: predict each sentence-internal break as being 150ms. We can make this model a bit smarter by building a "Mean" model: Instead of predicting 150ms, we can predict the mean value of breaks that we saw in the training data for that cross-validation fold. Third, we built a style-specific model as described before. Finally, we built a non-style-specific model, or a combined model: we combine the same cross-validation fold of all our corpora and train a combined CART model. The purpose of this combined model is to provide us with a reference performance of a model that is trained using method similar to our style-specific method, but still is not style-specific. The

hope is that style specific models will be better than the generic, combined model.

For each fold in the cross validation, we estimated the RMSE and Correlation number of our prediction using the four models at hand. We then averaged out the results over all cross-validation folds and looked at the average result for each corpus.

Table 2 shows the RMSE error of the four models on each corpus. Table 3 similarly shows the correlation of prediction. The RMSE and correlation values are on the prediction of the duration in the log-domain.

Table 2: RMSE of Predicted Duration (log-seconds domain)

Corpus	Festival	Mean	Combined	Style Specific
Europarl	0.4099	0.3858	0.6167	0.4186
ARCTIC	0.3897	0.3313	0.6858	0.3422
F2B	0.6778	0.4433	0.4794	0.4360
Obama	1.0456	0.7199	0.8685	0.7491
TATS	0.6736	0.5934	0.6021	0.5934
Emma	0.7072	0.6563	0.5834	0.5697

Table 3: Correlation of Predicted Duration (log-seconds domain)

Corpus	Festival	Mean	Combined	Style Specific
Europarl	0.0000	-0.0000	0.1939	0.0653
ARCTIC	0.0000	0.0000	0.2974	0.2174
F2B	0.0000	0.0000	0.1251	0.2770
Obama	0.0000	0.0000	0.0790	0.0868
TATS	-0.0000	0.0000	0.2276	0.1885
Emma	0.0000	0.0000	0.4634	0.5096

Looking at the RMSE, we see that the style specific model does generally perform better than the Festival model as well as the combined model. However, quite often, the mean model seems to get a lower RMSE. This is a bit surprising because it is a naive model, and moreover, the underlying distribution of breaks is not even normal. One possible explanation here is that since there is little training data on most of our corpora (one or two breaks in every utterance), our models are over-fitting to the training data across all cross-validation folds, leading to weak final model. If we look at the Emma corpus (our largest corpus), we see the results as we would expect: the Festival baseline does the worst, the mean-prediction does slightly better, followed by the combined model, and the style-specific model has the least error.

While RMSE is an important dimension to consider for evaluating our models, achieving the right speaking style means we should get a good correlation measure too. The mean model predicts a fixed value, and hence is not correlated at all with the actual duration. Our CART models can have a good correlation. The combined model typically gets better correlation numbers than the style specific model, even though it typically has higher error. On the largest (Emma) corpus, however, we see that the style specific model has higher correlation than the combined model, as we would expect.

We looked at the features that get used in the CART training. The type of punctuation that occurs before the break seems to make a good predictor for duration. The names of previous and next segments, as well as the part of speech features of adjacent words seem to be good predictors too. The number of words or stressed syllables in the previous phrase was not consistently a useful feature for pause duration. Analysis by [19] suggests that having more than 10 syllables in the preceding phrase strongly correlates with a long break. In our style corpora however, we did not notice such strong relationship between length of the current phrase and the break duration. Note however, that the prediction of where breaks should be inserted in the first place, does depend on the length of the current phrase.

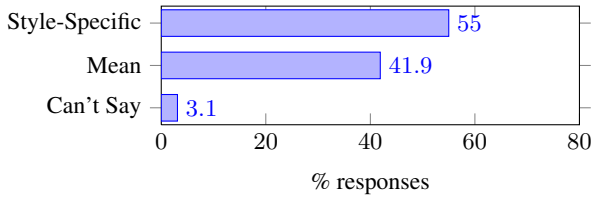
7. Subjective Tests

Objective results show that style specific duration is better than other models, on the Emma corpus. We ran subjective tests using this corpus to understand two aspects of duration modeling. First, we wanted to find out if people can even perceive differences in break duration for synthetic speech. Strictly speaking, we did not combine the pause duration into any other prosodic model (such as F0), and hence we wanted to investigate the impact of the duration alone, on perception. Secondly, if people can indeed perceive the difference between pause duration, we wanted to find out if they prefer to listen to synthesis that uses the style-specific duration model.

We ran two subjective comparisons. First, we compared the "Mean" model to the style specific model. We also compared the "Combined" model to the style specific model. We used a preference test for both these comparisons. We chose 25 utterances from our test data and synthesized them with the three models of phrase duration. In this case, we used Festival's default phrasing model to predict the location of breaks. We then created two tasks on Amazon Mechanical Turk to compare the two pairs of models. In each task, we presented the 25 utterances in random order. The two versions of each utterance were presented in random order, and were not labeled. Workers were asked to select the version that they preferred. We allowed up to 10 workers to do our tasks, and thus for each pair of comparison we have up to 250 data points of comparison. We filtered out responses by listeners that our automatic heuristics flagged as being spammers. The model that received the most votes by listeners can be considered to be the better one.

Figure 2 shows that the the style specific model performed better than the Mean model. This suggests that people perceive and prefer variability of pause duration in speech. Figure 3 shows that people could not tell the style-specific and combined models apart. Does this mean people can not tell apart subtle differences in variable duration? We plan to investigate this further.

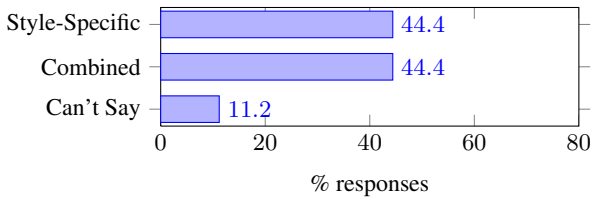
Figure 2: Subjective Result: Listener Preference for the Style-Specific model versus the Mean model



8. Conclusion and Future Work

In this paper, we looked at six different speech corpora of varying styles and sizes. Our analysis showed that the distribution of duration of breaks within sentences is quite different for each corpus. We showed that the distribution is not normal even in

Figure 3: Subjective Result: Listener Preference for the Style-Specific model versus the Combined model?



the log domain, but that we maybe able to model it as a Gaussian mixture.

We showed that a data driven method can be used to build duration prediction models for different styles. We compared these models to two naive models (Festival, Mean) and a non-style-specific model. Objectively, we found that the style specific model does better than the Festival and Combined models. Our results were stronger on the larger Emma corpus than on the other smaller corpora. We plan to investigate if this was simply because of the data size, or because of other dimensions on which the corpora differ (such as range of expression: Emma is the most free style among our corpora). If data scarcity is the underlying problem, using CART which splits data at each node may not be optimal and we could look at other regression techniques such as SVR. Our subjective results showed that people can perceive the difference in break duration methods, and that shows promise in exploring style-specific duration more thoroughly.

Until recently, speech synthesis has usually focused on synthesizing one utterance at a time. However, paragraph synthesis is gaining popularity in domains such as audio-book synthesis. If synthesis happens at paragraph levels or higher, we have to start caring not just about breaks within an utterance, but also breaks between utterances in a paragraph, and breaks at the ends of paragraphs. Modeling the duration of these breaks is tricky, since databases from which we can train them are not readily available. [20] have proposed a method with which large speech corpora could be aligned to their text, to automatically build a corpus for TTS voices that includes information at sentence and paragraph boundaries. Our next step is to construct such databases for available audio-books and model the duration of all phrase breaks occurring in speech. We would need to use cross-sentence features, and implement a support for them in the Festival Framework.

We also want to combine our previous work [2] on phrase prediction, and related prosody work such as [21] with this work on pause duration prediction and evaluate how much the combined methods take us closer to synthesizing with appropriate speech styles.

9. Acknowledgment

This work was supported by the Fundação de Ciência e Tecnologia through the CMU/Portugal Program, a joint program between the Portuguese Government and Carnegie Mellon University.

10. References

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling english prosody," in *Proceedings of 2nd International Conference on Spoken Language Processing*, Banff, Alberta, Canada, October 1992, pp. 867–870.

[2] A. Parlikar and A. W. Black, "A grammar based approach to style specific phrase prediction," in *Proceedings of Interspeech*, Florence, Italy, August 2011, pp. 2149–2152.

[3] F. Goldman-Eisler, "The distribution of pause durations in speech," *Language and Speech*, vol. 4, no. 4, pp. 232–237, 1961.

[4] R. Dhillon, "Using pause durations to discriminate between lexically ambiguous words and dialog acts in spontaneous speeech," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3190–3194, 2008.

[5] W. N. Campbell, "Multi-level timing in speech," Ph.D. dissertation, Sussex University, U.K. Department of Experimental Psychology., 1992.

[6] P. A. Barbosa and G. Bailly, *Progress in speech synthesis*. New York: Springer Verlag, 1997, ch. Generation of pauses within the z-score model, pp. 365–381.

[7] D. H. Klatt, "The KLATTALK text-to-speech conversion system," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, France, May 1982, pp. 1589–1592.

[8] A. W. Black and P. Taylor, "The festival speech synthesis system: system documentation," Human Communication Research Centre, University of Edinburgh, Tech. Rep., January 1997. [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival

[9] M. Schröder and J. Trouvain, "The german text-to-speech synthesis system mary: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[10] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Boston University, Tech. Rep., March 1995. [Online]. Available: http://ssli.ee.washington.edu/papers/radionews-tech.ps

[11] A. Parlikar and A. W. Black, "Data-driven phrasing for speech synthesis in low-resource languages," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012.

[12] J. Hirschberg, *Prosody: Theory and experiment*. Kluwer Academic Publishers, 2000, ch. A corpus-based approach to the study of speaking style, pp. 335–350.

[13] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of Machine Translation Summit*, Phuket, Thailand, September 2005, pp. 79–86.

[14] J. Kominek and A. W. Black, "CMU arctic databases for speech synthesis," in *Proceedings of the 5th Speech Synthesis Workshop*, Pittsburgh, Pennsylvania, June 2004, pp. 223–224.

[15] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[16] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, Toulouse, France, May 2006, pp. 853–856.

[17] E. Campione and J. Véronis, "A large-scale multilingual study of silent pause duration," in *Proceedings of the 1st International Conference on Speech Prosody*, Aix-en-Provence, France, April 2002, pp. 199–202.

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

[19] E. Zvonik, "Pausing and the temporal organization of phrases. an experimental study of read speech." Ph.D. dissertation, University College Dublin, National University of Ireland, November 2004.

[20] K. Prahallad, A. R. Toth, and A. W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *Proceedings of Interspeech*, Antwerp, Belgium, August 2007, pp. 2901–2904.

[21] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "A statistical phrase/accent model for intonation modeling," in *Proceedings of Interspeech*, Florence, Italy, August 2011, pp. 1813–1816.