

Correlating Text with Prosody

Mohamed Abou-Zleikha, Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics, University College Dublin, Dublin, Ireland

mohamed.abou-zleikha@ucdconnect.ie, julie.berndsen@ucd.ie

Abstract

The prediction of prosody from text information has long been recognised as a requirement for natural sounding speech synthesis. While an examination of the relationship between text information and prosody typically focuses on the role of accent, duration and phrasing both from a statistical and rule-based perspective, this paper investigates the correlation between the similarities calculated with respect to text information and those calculated with respect to prosody from an exemplar-based perspective. Two text features are examined, the syntactic tree and the dependency tree, along with two prosody features, pitch and intensity. The work in this paper investigates 1) the correlation between text information and prosody information 2) the conditional membership probability between text information and prosodic information, and 3) the effect of the number of exemplars on the conditional membership probability.

Index Terms: prosody prediction, syntactic tree, dependency tree, prosody text correlation, prosody text similarity

1. Introduction

Prosody modelling is an important factor in speech synthesis where pitch and intensity contours play a demonstrable role in the intelligibility and naturalness of synthesised speech. Several studies on speech synthesis have been undertaken in order to predict such contour information from text. Some of these studies investigate the relationship between text information and prosody using rule-based systems, others use statistical and exemplar-based methodologies [1] [2] [3].

The work in this paper investigates the relationship between the text information and prosodic information from an exemplar-based perspective. The exemplar-based model aims to capture the detailed exemplar memory in which exemplars are stored with rich information. When a new input is presented, it is compared against the exemplars, and the closest exemplar (or set of exemplars) found to the new input is selected. In such models, the criteria for choosing an exemplar from the exemplars cloud play an important role in the model performance. These criteria depend on the text features of the input.

The purpose of the work presented in this paper is to study the relationship between text information (syntactic and dependency information) and prosodic information (pitch model, pitch contour, and intensity contour) from the exemplar-based point of view. The study focuses on finding answers to three main questions that are important for exemplar-based prosody prediction:

1. What is the relationship between similarity calculated with respect to text information (syntactic and dependency trees) and similarity calculated with respect to prosody (pitch and intensity contours), i.e. is there a correlation between them, and if there is a correlation, is it significant?

2. What is the probability of a sample utterance chosen from the set which represents the closest $\alpha\%$ to an element i on the text level being also an element of the set which represents the closest $\beta\%$ to that element i on the prosody level (i.e. the conditional membership probability)?
3. What is the impact of the number of exemplars on the conditional membership probability?

The conditional membership probability will answer the question as to whether for a new input, which does not exist in the corpus and for which only text information is available, an exemplar with similar text information can be found with associated prosodic information that can be adapted to serve as an appropriate prosody model for this input.

2. Previous work

The relationship between prosody of speech and syntactic structure of an utterance has been considered by several studies. The purpose of most studies is to improve the naturalness of synthesised speech. Traditionally, the syntactic tree, the dependency tree and part of speech features (POS) are used as potential text features. Some studies assume the existence of a relationship between text information and prosody [4]; others focus on defining the relationship between the syntactic or dependency tree and pitch accent, duration and phrasing from a statistical and rule-based point of view [5] [6] [7] [8] [9] [10] [11]. Nevertheless, accent, phrasing and duration are not the only important features in prosody. The global contours of pitch and intensity play also important roles in the prosody of synthesised speech. Analysing study is done on the relation between the dependency graph and prosodic phrasing and prominence [5]. In [11] the mapping between dependency trees prosodic phrasing and prominence has investigated to predict the prosody information in order to generate prediction rules. In [12] the same approach is used but by mapping to the syntactic tree.

Research on the similarity between the syntactic tree and timing tree has been conducted to define the mapping between two trees [6], and the mapping between the syntactic tree and prosody tree using tree edit distance function has been used to predict prosody in [4]. The relationship between intonational phrasing and syntactic structure in sentence production has also been investigated on the accent and phrasing boundary levels [7] [8] [10]; and the relationship between POS and accent was also a research question to find the mapping between these two features [9].

Some approaches have considered the relationship between the syntactic/dependency trees, pitch and intensity contours from an exemplar-based point of view [1], but none of these studies have used the notion of conditional membership probability.

3. Description of the Data

The CMU Arctic speech corpus is used for the study done in this paper. The corpus contains 1132 utterances spoken by a US English male speaker.

For each utterance, two types of information are extracted:

1. Information about the text that includes
 - Syntactic tree (ST): which represents the syntactic structure of an utterance according to the language grammar
 - Dependency tree (DEP): which represents the grammatical dependencies between words in the utterance.
2. Information about the prosody, which contains:
 - (a) Pitch contour with its velocity and acceleration
 - (b) Pitch contour represented using a pitch model
 - (c) Statistical parameters of the pitch contour (max, min, mean, standard deviation)
 - (d) Intensity contour with its velocity and acceleration
 - (e) Statistical parameters of the intensity contour (max, min, mean, standard deviation)

The syntactic and dependency trees are extracted using the Stanford parser [13] [14]. A distance matrix is calculated for each type of information and the model used in (b) has been presented in [15]. For the text information, two distance functions are used to calculate the distance matrix: the tree edit distance function (TED) [16] which represents the number of insertion, deletion and substitution operations required to transform the tree representation of one text to the tree representation of another, and tree kernel function distance (TKD) [17] which depends on calculating the existence of all subtrees of the first tree in the second one and vice versa. For the prosodic information, dynamic time warping (DTW) with Euclidean distance is used to calculate the distance matrix for (a) and (d) above. DTW with a special distance function for the model [15] is used for (b), and the Euclidean distance function is used for (c) and (e).

In the rest of this paper, the term ST-TED will be used for the distance matrix of the syntactic tree calculated according to the tree edit distance function, the term ST-TKD for the distance matrix of the syntactic tree calculated according to the tree kernel function, the term DEP-TED for the dependency tree calculated according to the tree edit distance function and the term DEP-TKD for the dependency tree calculated according to the tree kernel function.

4. Experiments

The experiments are performed to address the questions posed in section 1.

4.1. Similarity Correlation

In order to answer the first question, it is necessary to calculate the correlation between each pair of the distance matrices and to determine the correlation between each pair of columns, where each column represents the distance between the element at that index and other elements of the corpus. Given matrix X and matrix Y , the purpose of the first step is to calculate the correlations between each pair of columns (X_i, Y_i); e.g X_i is a ST-TED vector of example i , Y_i is the pitch model distance

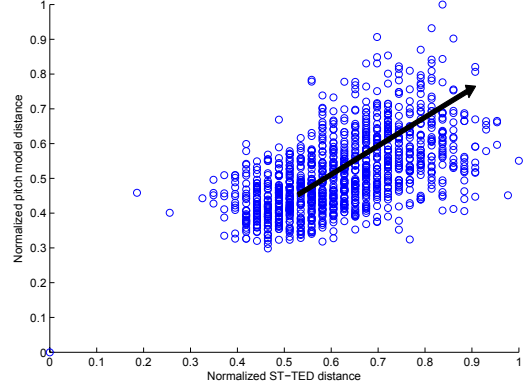


Figure 1: Scatter points of positive correlation (0.56) of two columns from normalised pitch model distance matrix and normalised ST-TED matrix; both columns have the same index.

with a correlation 0.56. A high positive value for the correlation coefficient indicates that the correlation between the two features is high and has a positive direction, which means that the similarity values between the features increase in the same direction as illustrated in Figure 1.

The second step is to take the average value of these correlations. Table 1 presents the result of the average correlation. To check if the correlation is statistically significant, the p value is calculated for each of the correlations and then taking the average as shown in Table 2. The correlation values are merely the threshold that indicates which features should be taken into consideration in the next stage.

Table 1: Average correlation values between the studied text features and prosody features distance matrices.

	ST-TED	DEP-TED	ST-TKD	DEP-TKD
Pitch model	0.476	0.475	-0.054	0.059
Pitch contour	0.336	0.306	0.006	0.062
Pitch statistical	0.023	0.008	0.011	0.022
Intensity contour	0.377	0.344	0.018	0.075
Intensity statistical	0.0006	-0.0009	0.052	0.051

Table 2: Average p -values for the correlation between the studied text features and prosody features distance matrices.

	ST-TED	DEP-TED	ST-TKD	DEP-TKD
Pitch model	3.3e-07	0.00057	0.148	0.144
Pitch contour	0.036	0.032	0.137	0.15
Pitch statistical	0.183	0.227	0.385	0.425
Intensity contour	0.023	0.0262	0.141	0.121
Intensity statistical	0.359	0.411	0.212	0.191

The results in Table 1 and Table 2 show significant correlation between the similarity of the syntactic tree and the dependency tree using tree edit distance function with the pitch model, the pitch contour and the intensity contour. However, results do not indicate a significant correlation with the statistical parameters of the pitch and the intensity contours. The tree kernel function similarity does not show a significant correlation with any of the prosody information. The rest of this paper focuses only on the matrices that show a significant correlation.

Once the significantly correlated matrices are identified, the correlations between the combinations of these matrices are calculated:

1. ST-TED + DEP-TED
2. Pitch model + Intensity contour.
3. Pitch contour + Intensity contour.

These combinations are calculated by summing the normalised distance of each matrix. Table 3 presents the correlation coefficients and p-values for the combination matrices.

Table 3: Correlation coefficients and p-values for the combination matrices

	ST-TED +DEP-TED	
	correlation coefficient	p-value
Pitch model + Intensity contour	0.510	0.0039
Pitch contour + Intensity contour	0.404	0.0235

4.2. Conditional Membership Probability

The purpose of this experiment is to answer the question posed in the introduction section above, namely, what is the probability of a sample utterance chosen from the set which represents the closest $\alpha\%$ to an element i on the text level being also an elements of the set which represents the closest $\beta\%$ to that element i on the prosody level. This requires the calculation of the conditional membership probability.

Given:

- The prosody information $X = [x_1 \dots x_n]$; where $x_i = [x_{i1} \dots x_{in}]$ is the distance vector between example i and the rest of the data according to the defined prosody distance function, x_i a sorted distance vector to the example i according to the prosody information.
- The text information $Y = [y_1 \dots y_n]$; where $y_i = [y_{i1} \dots y_{in}]$ is the distance vector between example i and the rest of the data according to the defined text distance function, y_i a sorted distance vector to the example i according to the text information.

If the closest $\alpha\%$ from x_i is chosen and called $x\alpha_i$ and the closest $\beta\%$ from y_i is chosen and called $y\beta_i$, then

$$P(z \in x\alpha_i | z \in y\beta_i) = \frac{P(z \in x\alpha_i \cap z \in y\beta_i)}{P(z \in y\beta_i)} \quad (1)$$

The experiment was carried out on the matrices which show a significant correlation for two text information distance matrices: ST-TED, and DEP-TED; and three prosody matrices; pitch model distance matrix, pitch contour distance matrix and intensity contour distance matrix. Different values for α and β were used to calculate the probability for each pair of matrices. Table 4 illustrates the conditional membership probability between ST-TED and pitch model distance matrix.

The results from Table 4 shows that increasing the posterior membership data increases the conditional membership probability, but this also leads to a decrease in the accuracy of selection. On the other hand, increasing the prior membership data has no large effect on the probability value, but this leads to a reduction in the size of the intersection between the prior and posterior data and a decrease in the accuracy of selection as well. A trade-off is needed between the membership probability and the percentage of selected data which represents the

Table 4: Conditional membership probabilities between ST-TED closest data and pitch model closest data

		ST-TED closest data (prior)				
		10%	20%	30%	40%	50%
Pitch model	10%	0.28	0.25	0.23	0.21	0.20
	20%	0.46	0.43	0.40	0.37	0.34
closest data	30%	0.60	0.58	0.54	0.51	0.47
	40%	0.72	0.70	0.66	0.63	0.60
(posterior)	50%	0.81	0.79	0.77	0.74	0.71

accuracy of the similarity. From Table 4 we can see that choosing 10% for the text information and 30% from the prosody information gives good accuracy. The same experiment was carried out for the other distance matrices. Table 5 illustrates the probability of picking an exemplar from 10% closest exemplars to the input for text information, and finding it in the 30% of closest exemplars in prosody information, where the first three rows represent the conditional membership probability calculated using independent models; the fourth and fifth represent the conditional membership probability using the combinations of similarity matrices, and the last two rows represent the conditional membership probability of an exemplar belonging to two similarity matrices (\wedge).

Table 5: The conditional membership probability for 10% closest exemplars in text information, and 30% of closest exemplars in prosody information.

	ST-TED	DEP-TED	ST-TED+ DEP-TED
Pitch model	0.60	0.60	0.64
Pitch contour	0.51	0.49	0.52
Intensity contour	0.54	0.52	0.55
Pitch model + Intensity contour	0.61	0.59	0.63
Pitch contour + Intensity contour	0.55	0.53	0.56
Pitch model \wedge Intensity contour	0.38	0.36	0.40
Pitch contour \wedge Intensity contour	0.35	0.33	0.35

4.3. The effect of the number of exemplars

To determine the effect of the number of exemplars on the conditional membership probability, different numbers of exemplars are used to calculate the conditional membership probability using 30% closest exemplars to an external example on the prosody level and the 10% of closest exemplars on the text level. Figure 2 shows the conditional membership probability for different matrices along with different numbers of exemplars.

The results show a linear relationship between the conditional membership probability and the number of exemplars in the data. Increasing the number of exemplars results in an increase in the membership probability but it also leads to an increase in computational time and causes a reduction in the accuracy of selection. For example, increasing the exemplar size from 100 to 1000 yields an increase in the size of the data set i.e. the best 10% of 100 exemplar yields a set of 10 exemplars while the best 10% of 1000 exemplars yields a set of 100 exemplars.

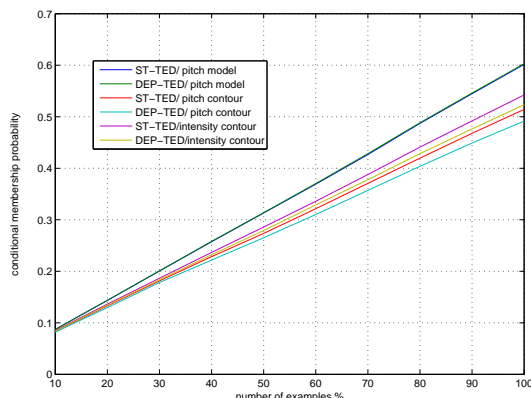


Figure 2: Conditional membership probability per number of exemplars in percentage scale.

5. Discussion and Conclusion

This paper presented a set of experiments to examine the correlation between the text information and prosody information. Studying of the relationship between these two types of information and defining a probability measure for this relationship provide the basis for identifying criteria for exemplar-based prosody modelling [15]. Three experiments were performed to achieve this goal. The first experiment showed a significant correlation between three features of prosody information; pitch model, pitch contour and intensity contour, and two features from the text information; the syntactic and dependency tree. The experiments showed a strong dependency between the distance function and the correlation. In the second experiment, the results indicated good conditional membership probabilities for the correlated data, about 60% for pitch model, 54% for the intensity contour and about 40% for the $pitchmodel \wedge intensitycontour$ when 10% from the text information and 30% from the prosodic information are chosen. The third experiment showed a linear relationship between the conditional membership probabilities and the number of used exemplars, the more data it is used the higher membership probability it is observed, but this increases the computational time and of course decreases selection accuracy. It is difficult to compare these results with other approaches due to the fact that it uses different features and methodologies to those employed elsewhere; a comparison with other approaches will be possible when the results are incorporated into the speech synthesis system.

The results of these experiments answered the main questions on the prosody and text levels and provide selection criteria for exemplar-based prosody modelling. The importance of this step is that it overcomes the shortcomings of traditional exemplar selection, where the probability that the best exemplar is chosen is very low. The next step involves integrating the exemplar-based prosody model into unit selection speech synthesis using the correlation between the text information and prosodic information as a criterion to select an appropriate exemplar from the corpus. Further work is required to investigate the effect of using an intensity model instead of the intensity contour on the correlation and also the relationship between the duration and text information.

6. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Center for Next Generation Localisation (www.cngli.ie) at University College Dublin, Ireland. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

7. References

- [1] T. Yoon, "A predictive model of prosody through grammatical interface: A computational approach," *Doctoral dissertation, University of Illinois at Urbana-Champaign*, 2007.
- [2] V. Sridhar, A. Nenkova, S. Narayanan, and D. Jurafsky, "Detecting prominence in conversational speech: pitch accent, givenness and focus," *Proceedings of Speech Prosody, Campinas, Brazil*, pp. 380–388, 2008.
- [3] P. Taylor, S. King, S. Isard, and H. Wright, "Intonation and dialog context as constraints for speech recognition," *Language and Speech*, vol. 41, no. 3-4, p. 493, 1998.
- [4] L. Blin and M. Edgington, "Prosody prediction using a tree-structure similarity metric," in *Sixth International Conference on Spoken Language Processing*, 2000.
- [5] A. Lindstrom, I. Bretan, and M. Ljungqvist, "Prosody generation in text-to-speech conversion using dependency graphs," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3. IEEE, 1996, pp. 1341–1344.
- [6] D. Gibbon, "Corpus-based syntax-prosody tree matching," in *Eighth European Conference on Speech Communication and Technology*, 2003.
- [7] D. Watson and E. Gibson, "The relationship between intonational phrasing and syntactic structure in language production," *Language and Cognitive Processes*, vol. 19, no. 6, pp. 713–755, 2004.
- [8] A. Cohen, "A survey of machine learning methods for predicting prosody in radio speech," *Master's thesis, University of Illinois at Urbana-Champaign*, 2004.
- [9] S. Arnfield, "Prosody and syntax in corpus based analysis of spoken english," *PhD Thesis, School of Computer Studies and Psychology, University of Leeds*, 1994.
- [10] A. Black and P. Taylor, "Assigning intonation elements and prosodic phrasing for english speech synthesis from high level linguistic input," in *Third International Conference on Spoken Language Processing (ICSLP 94), Yokohama, Japan, September 18-22, 1994. volume 2*, pp. 715–718., 1994.
- [11] J. Hirschberg and O. Rambow, "Learning prosodic features using a tree representation," *Seventh European Conference on Speech Communication and Technology*, 2001.
- [12] I. Koutny, G. Olaszy, and P. Olaszy, "Prosody prediction from text in hungarian and its realization in its conversion," *International Journal of Speech Technology*, vol. 3, no. 3, pp. 187–200, 2000.
- [13] D. Klein and C. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 423–430.
- [14] M. De Marneffe, B. MacCartney, and C. Manning, "Generating typed dependency parses from phrase structure parses," in *Proceedings of LREC*, vol. 6, 2006, pp. 449–454.
- [15] M. Abou-Zleikha, P. Cahill, and J. Carson-Berndsen, "An automatic pitch model with distance function," in *Proceedings of the the Seventh ISCA Speech Synthesis Workshop*, 2010.
- [16] E. Demaine, S. Mozes, B. Rossman, and O. Weimann, "An optimal decomposition algorithm for tree edit distance," *Automata, languages and programming*, pp. 146–157, 2007.
- [17] M. Collins and N. Duffy, "New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 263–270.