interspeech 2008 incorporating SST 08

Including Pitch Accent Optionality in Unit Selection Text-to-Speech Synthesis

Leonardo Badino, Robert A.J. Clark, Volker Strom

Centre for Speech Technology Research
University of Edinburgh, Edinburgh, UK
l.badino@sms.ed.ac.uk

Abstract

A significant variability in pitch accent placement is found when comparing the patterns of prosodic prominence realized by different English speakers reading the same sentences.
In this paper we describe a simple approach to incorporate this variability to synthesize prosodic prominence in unit selection text-to-speech synthesis.
The main motivation of our approach is that by taking into account the variability of accent placements we enlarge the set of prosodically acceptable speech units, thus increasing the chances of selecting a good quality sequence of units, both in prosodic and segmental terms.
Results on a large scale perceptual test show the benefits of our approach and indicate directions for further improvements.
Index Terms: speech synthesis, unit selection, prosodic prominence, pitch accents

10.21437/Interspeech.2008-549

1. Introduction

This paper describes a novel approach to model prosodic prominence and exploit its intrinsic variability (observed in prosodically annotated speech corpora) in order to improve both the prosodic and segmental speech quality of a unit selection text-to-speech (TTS) synthesis system.
We will focus on pitch accents as they are widely regarded as prosodic events that signal prominence.
Although the literature on prosodic prominence reports several works on pitch accent prediction from explanatory textual features, it is only recently that the impact of pitch accent detection in prosodic modeling for speech synthesis has been investigated. In [1] a pitch accent predictor is used to annotate the speech database (containing both neutral and expressive prosodic style speech) and the sentences to be synthesized. Besides, a distinction between standard and emphatic (particularly strong) accents is made. Both pitch accents and emphatic accents placements are then treated as specifications for the target cost function. A large scale perceptual test was carried out to investigate the benefits arising from taking into account standard and emphatic accents. Although the empirical evidence showed a significant advantage from using standard accents, emphatic accents turned out to be more determinant than standard accents and the best results were achieved when both type of accents were specified in the target cost function.
It is worthwhile noting that while the standard accents were automatically annotated, emphatic accents were already marked up by the authors of the sentences in the form of capitalized words and during the speech database recording the speaker was required to emphasize capitalized words.
The work we are going to describe here is strictly related to [1]. The main difference is that here we focus on standard pitch accents only and we introduce the concept of pitch accent optionality which directly stems from the variability of prosody.
When different speakers are required to read the same sentence, prosodic diversity is usually observed. This diversity involves pitch accents placement and shape, phrase breaks location and other aspects of prosody, and it is even observed when it is a single speaker who reads the same sentence at two different times [2]. The variability observed in pitch accents and phrase breaks placement has suggested a distinction between optional and compulsory prosodic events, a distinction that, in turn, has been used to reformulate the evaluation of automatic predictors ([3],[4],[5]).
The core idea of the present work is that of associating to each accent (and to each no-accent) placement prediction its supposed degree of optionality, expressed, as we show and motivate later, as the "uncertainty" of the accent predictor. The motivation to to do this is that by incorporating the optionality of the accent we enlarge the set of prosodically acceptable speech units, and so increase the chances of selecting a good quality sequence of units, both in prosodic and segmental terms. For example, let us suppose that a "highly optional" has been predicted for a given syllable (of an input sentence to a TTS system). Because of the high optionality of that accent, a deaccented syllable would be probably equally acceptable and so we can allow the unit selection module to select either accent-bearing or no-accent bearing speech units. Doing so we loosen the prosodic constraints without worsening our prosodic model and consequently we increase the number of available candidate units.
The advantages of incorporating prosodic variability to improve the quality of unit selection speech synthesis have already been shown in some recent works([6],[2],[7] and [8] among them). For example in [7] different intonation contours are generated via unit selection using prosodic target cost features such as position of the syllable in the intonational group, accents(but no optionality taken into account), etc... The generated contours become then target contours for the standard unit selection phase and the sequence of speech units having the lowest combined (prosodic plus segmental) cost is chosen. Both objective and perceptual test show the clear benefits of using more than one target intonation contours.
Although exploiting prosodic variability is not a novel idea, we are now proposing of taking into account the optionality of some prosodic events (such as pitch accents or phrase breaks) in a way that the automatic prediction of such events is no longer a stand-alone step preceding the unit selection phase, but becomes an integral part of the unit selection process itself.

Copyright © 2008 ISCA
Accepted after peer review of full paper

	f1a	f2b	f3a	m1b	m2b	m3b
may	N	A	A	A	A	A
be	N	N	N	N	N	N
the	N	N	N	N	N	N
most	N	A	N	A	N	A

Figure 1: *An example extracted from the BURN corpus. A and N stand for accent and no-accent respectively.*

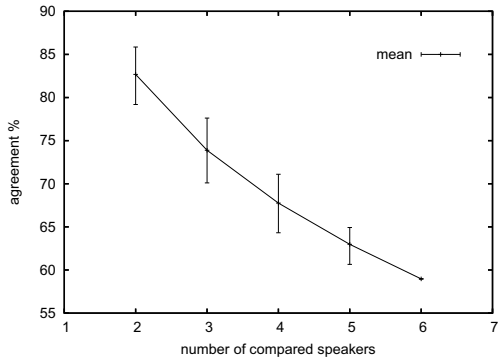

Figure 2: *Speakers agreement in pitch accent placement. The mean line represents the mean disagreement value.*

2. Pitch accent optionality

2.1. Disagreement among speakers

In [5] we analyzed a section of the Boston University Radio News (BURN) corpus [9] containing the speech of six different speakers (three females: f1a, f2b, f3a, and three males: m1b, m2b, m3b) reading the same text. All data have been prosodically labeled using the ToBI annotation conventions. We used this annotation only to see if a pitch accent occurred or not (see Figure 1).

Figure 2 shows the percentages of intra-speaker agreement for each combination of speakers and the agreement mean, with respect to the number of speakers involved, on a text of 1662 words. The error bars range from the lowest to the highest agreement percentage, given a certain number of speakers. For example, given a number of two speakers, there are 15 possible combinations of speakers. Among them the pair with the lowest agreement (79.19%) is f1a-m2b, whereas the highest agreement (85.86%) occurs in m1b-m3b.

2.2. An alternative definition of optionality

In pitch accent literature the optionality of a pitch accent has always been considered as a simple binary variable and when analysing multi-speaker data such as the BURN corpus, a pitch accent is considered optional if m out of n speakers, with $n =$ *total number of speakers* , $0 < m < n$ (and usually $m = 1$), realize that accent.
Such a definition of optionality has the disadvantage of being strongly dependent on n; indeed, by simply adding a speaker to the multi-speaker corpus, an accent can suddenly change its status from compulsory to optional. For this reason it seems more appropriate to consider different degrees of optionality, that is, defining optionality as a gradient variable. In [5] we formulated a new definition of optionality by associating an emission source to each word token. Each source can emit two symbols, one when the token is accented and one when it is not. The number of emissions is equal to the number of speakers and each emission is independent from the others.
From Information Theory we know that the entropy of such a source is:

$$H = -\log(P(A))P(A) - \log(P(B))P(B) \qquad (1)$$

where $P(A)$ is the probability that the source emits an accent and $P(N)$ that it does not. The entropy says how much information we need (or more informally, how many questions we have to ask) to correctly predict the next symbol that will be emitted by the source. If the source has always emitted the same symbol than its entropy will be 0, whereas if the number of emissions of both symbols is equal then the entropy value will be 1. In all the other cases (and if the number of emissions is higher than 2) the entropy value will be less than 1 and more than 0.
Since H is a measure of the source uncertainty we can use it as a measure of optionality as well. The higher the source uncertainty the higher the symbols optionality.

word	f1a	f2b	f3a	m1b	m2b	m3b	H
w_{i-2}	N	A	A	A	A	A	0.9182
w_{i-1}	N	N	N	N	N	N	0
w_i	A	A	A	A	A	A	0
w_{i+1}	N	A	N	A	N	A	1
w_{i+2}	N	N	N	A	N	A	0.6500

Table 1: *Examples of H values. The entropy value H is given by equation 1. In a speech corpus of 6 speakers reading the same sentences, there are 4 possible values of optionality: 0, 0.6500, 0.9182 and 1.*

3. Including the accent predictor uncertainty in the target cost function

In order to incorporate pitch accent optionality in our TTS system we could build an accent optionality predictor trained on the multi-speaker part of the BURN corpus, having the task of correctly predicting the H value per each word token, which is given by 1 (see example in Table 1). Then we could associate the predicted optionality value to the accent predictor's prediction (accent/no-accent) and use it to tune the cost associated to the pitch accent target cost feature.
Let us consider the standard target cost function for a unit selection speech synthesis system:

$$T(s_t, u_t) = \sum_{s=1}^{F} w_f(T_f(s_t[f], u_t[f])) \qquad (2)$$

where $s_t[f]$ and $u_t[f]$ are the values for the feature f of the target unit and the candidate unit respectively, T_f is the function evaluating the distance between $s_t[f]$ and $u_t[f]$, and w_f is the weight of the feature f.
Instead of using a standard T_f for the pitch accent feature, that is a T_f that returns 0 when $s_t[f]$ and $u_t[f]$ have the same value and 1 when they have two opposite values, we could introduce

the following T_f:

$$T_f = \begin{cases} 0 & \textit{if } s_t[f] \textit{ and } u_t[f] \textit{ are equal} \\ 1 - H(w_i) & \textit{otherwise} \end{cases} \quad (3)$$

where $H(w_i)$ is the accent optionality associated to the word containing the target unit s_t.
When $s_t[f]$ and $u_t[f]$ are different and $H(w_i)$ is close to 0, i.e. the pitch event (accent or no-accent) is highly compulsory, then T_f returns a value very similar to the standard T_f, whereas when the pitch event is highly optional $H(w_i)$ is close to 1 and the cost associated to the pitch accent feature is very low: it does not really matter if the speech unit comes from an accented syllable or not.
What we actually did in our implementation was using a $H(w_i)$ that is not a predicted value of optionality but the value of uncertainty of our pitch accent predictor for that given word. Using a pitch accent predictor that for each predicted pitch event e_i gives the conditional probability $P(e_i|F_i)$, where F_i is a window of explanatory features centered at position i, the uncertainty of its prediction simply is:

$$\begin{aligned} H_p(w_i) = \quad & -\log(P(e_i|F_i))P(e_i|F_i) \\ & -\log(1 - P(e_i|F_i))(1 - P(e_i|F_i)) \end{aligned} \quad (4)$$

The main motivation to use $H_p(w_i)$ instead of $H(w_i)$ is that what we actually need to know is how "sure" is the accent predictor about the information it passes to the target cost function. For example, since we deal with not perfect predictors, it may happen that the accent predictor assigns a pitch accent to a word with $P(e_i) = 0.55$ and the optionality predictor (the one predicting the H value of equation 1) says the accent is compulsory ($H = 0$). In that case what it is important for the unit selection module is that the accent predictor is quite unsure about its prediction and so it does not make a lot of sense forcing the selection module to select accented units, independently from the optionality predictor.
Looking from another perspective, the advantage of using $H_p(w_i)$ is that it is correlated to both the optionality of a pitch accent and the inaccuracy of the accent predictor. In fact our accent predictor does not achieve a 100% accuracy rate because: 1) the set of explanatory features we use is not enough (inaccuracy of the prediction model); 2) prosodic variability occurring within a single speaker's speech. As a consequence $P(e_i|F_i)$ is affected by both factors.
However, the assumption that the H_p of a accent predictor is correlated to pitch accent optionality is not necessarily true and depends on the accent predictor. Making an extreme example, let us assume of having an accent predictor trained on a single explanatory feature having two different values: $true$ if the first letter of a word is a d, and $false$ otherwise. In that case is easy to see that H_p is only due to the inaccuracy of the prediction model. Instead, if the explanatory features are "good enough" the optionality observed with respect to those features should be strongly correlated to the real optionality.
The accent predictor we used is a Classification and Regression Tree (CART) (from [10]), trained on the f2b section of the BURN corpus, that for each predicted pitch event gives the conditional probability $P(e_i|F_i)$, where F_i is a window of explanatory features centered at position i and covering five words. The explanatory features are logarithms of the unigram and of the bigram of the probability of a word (computed on the Herald news (1998-2002) corpus), and Part-of-Speech (given by the MXPOST tagger [11]) of a word. The results achieved by our predictor are comparable with that of state-of-the art pitch accent predictors (see [5] for details).
To see if the uncertainty of our accent predictor was related to optionality we built an optionality predictor as in [5] and used the H_p of our accent predictor as an explanatory feature. It turned out that the higher H_p was the more optional the accent was.
Note that in this case, our accent predictor was trained on single speaker data, while the optionality predictor was trained on multi-speaker data, so this result, together with other results in [5] showing that accent predictors trained on single speaker data and accent predictors trained on multi-speaker data achieve very close accuracy rates on several test data, seems to confirm that, at least with respect to our prediction model features, the variability occurring within a single speaker and the variability occurring among speakers are two overlapping phenomena.

In order to test if including H_p in the target cost function does produce any benefit we compared a TTS system (henceforth THP) having the modified T_f of equation 3 (with H_p instead of H) with one (henceforth TC) having the standard T_f. As we mentioned above, we expected THP to produce on average a better quality speech because of the looser prosodic constraints that enlarge the search space of the the unit selection module without worsening the prosodic model. Nevertheless our expectations are based on an approximation we have made so far: the placement of a pitch accent does not depend on the placements of the accent preceding it. If this approximation is completely wrong then our definition of optionality (and that one of previous works) is wrong as well, since it considers each pitch accent as a stand-alone event.

4. Test design

To create a set of utterances for evaluation we run our pitch accent predictor on a subsection of the BURN corpus and of the Herald news and on sentences no longer than 20 words. The pitch event prediction and its H_p value were assigned to each word and the sentences were ranked according to the average value of H_p per each word from the highest to lowest. The first 15 sentences having the highest rank and producing audible differences between the two systems were chosen for the test.
Each sentence was synthesized using both TTS systems so 15 pairs of utterances were generated. Each pair was presented to each participant twice but in reversed versions (i.e. THP-TC and TC-THP) so each participant listened to a total of 30 pairs whose order was randomized per each participant. The experiments were carried out through a web browser and, for each pair, the participants could express a preference for one of the two utterances, or no preference.
46 subjects were recruited, all of them are native English speakers. The tests lasted approximately 20 minutes each.

5. Results

The overall results are shown in Table 2. We computed the number of preferences for the three options (THP, TC and No preference) on the overall set of pairs, and on the set only containing pairs where the subject's choices were consistent, that is were the subject chose the same option in both pairs. We then run two different kinds of two-sided Binomial tests: one where all the preferences for the "No preference" option were excluded and one where they were split into two equal halves and one half was summed to the THP preferences and the other one to the TP preferences.

	THP	TC	No-preference	p-value_1	p-value_2
All pairs	587	439	354	$p < 0.00001$	$p = 0.00007$
Consistent preferences only	191	103	78	$p < 0.00001$	$p < 0.00001$

Table 2: *Comparison between THP and TC. In the* All pairs *row the comparison is made on all the pairs* $(30 * 46)$ *presented in the experiments. In the* Consistent preferences only *row the comparison is made only on the pairs where the preferences of the subjects where consistent. The first three columns report the number of preferences for the three options. The p-values are from two-sided Binomial tests. The p-value_1 is computed by excluding the* No preference *choices from the overall set of choices, while the p-value_2 is computed by splitting the* No preference *set in two halves and summing one half to the THP preferences and one half to the TC preferences.*

Sentence ID	THP	TC	No-preference
Sentence 2	23	53	16
Sentence 3	78	8	6

Table 3: *Sentences with highest number of preferences for THP and TC respectively*

All conditions and tests show a statistical significant preference for the TTS that embodies H_p, with p-values far below 0.001. Looking at each single sentence the difference between the two systems is less evident since for only 5 sentences out of 15 there is a significant (p-value_2 < 0.01) preference for THP, whereas for 4 sentences the significant preference is for system TC, and for the remaining 6 sentences there is no significant preference for neither of the two systems. Despite of this small difference between THP and TC the overall results show a significant preference for THP because, when significant, the preferences for THP are more definite than the preferences for TC (see Table 3).

Listening to the test utterances we noticed that in a couple of cases what we perceived was the opposite from what we would have expected: where the value of H_p for a given word was high, that word was strongly accented by the THP system and not accented or slightly accented by the TC system. We believe this behaviour is mainly due to: 1) the intrinsic "instability" of the unit selection technique; 2) the inaccuracy of the pitch accent annotation in the speech database. The annotation inaccuracy is a consequence of the fact that the voice of the speech database and that on which the pitch accent predictor was trained are different, and we have seen in Section 2 how much different the pitch accent sequences of two speakers can be. As a consequence we believe that a better annotation of the speech database, may be enhanced using acoustic explanatory features, may reduce unexpected outputs.

6. Conclusions

We have proposed a method to include prosodic prominence variability in a unit selection TTS system that models prosodic prominence by using pitch accents as specifications for the target cost function.

Results from a large scale perceptual experiment support our working hypothesis: including prosodic prominence variability does not worsen the prosodic model and increases the chances of selecting good quality sequences of speech units.

Finally our listening tests also point out the necessity for a better pitch accents annotation of the speech database in order to achieve a more coherent realization of prosodic prominence in unit selection TTS synthesis.

7. Acknowledgments

This research was funded by the Institute for Communicating and Collaborative Systems (University of Edinburgh), the College Science Fund, the Chalmers Fund and the Colin and Ethel Gordon Fund.

8. References

[1] V. Strom, A. Nenkova, R. Clark, Y. Vasquez-Alvarez, J. Brenier, S. King, D. Jurafsky, "Modelling Prominence and Emphasis Improves Unit-Selection Synthesis" in Proc. Interspeech 2007, Antwerp, Belgium, 2007.

[2] M. Chu, Y. Zhao, E. Chang (2006) "Modeling stylized invariance and local variability of prosody in text-to-speech synthesis" Speech Communication 48, 716-726.

[3] E. Marsi (2004). "Optionality in Evaluating Prosody Prediction". Proc. Of 5th ISCA Speech Synthesis Research Workshop, Pittsburgh, USA.

[4] J. Yuan, J. M. Brenier, D. Jurafsky (2005). "Pitch Accent Prediction: Effects of Genre and Speaker". Proc. Interspeech 2005, Lisboa, Portugal.

[5] L. Badino, R A. J. Clark (2007). "Issues of optionality in pitch accent placement". Proc. 6th ISCA Speech Synthesis Workshop, Bonn, Germany.

[6] J. Bulyko and M. Ostendorf, "Joint prosody prediction and unit selection for concatenative speech synthesis", in Proc. of ICASSP 2001, Salt Lake City, USA, 2001.

[7] F. Campillo, R. Banga (2006). "A method for combining intonation modeling and speech unit selection in corpus-based speech synthesis systems". Speech Communication 48, 941-956.

[8] R.A.J. Clark, S. King (2006). "Joint Prosodic and Segmental Unit Selection Speech Synthesis". Proc. Interspeech 2006, Pittsburgh, USA.

[9] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis", Computer Speech and Language, 10:155-185, 1996.

[10] P.Taylor, R. Caley, A.W. Black, and S.King, "Edinburgh Speech Tools Library" System Documentation Edition 1.2, for 1.2.0 15th June 1999.

[11] "A Maximum Entropy Part-of-Speech tagger" in Proc. of the Empirical Methods in natural Language Processing Conferece, University of Pennsylvania, 1996.