Towards Hierarchical Prosodic Prominence Generation in TTS Synthesis

Badino, L., Clark, R. A. J., & Wester, M. (2012). Towards Hierarchical Prosodic Prominence Generation in TTS Synthesis. In INTERSPEECH 2012 13th Annual Conference of the International Speech Communication Association. (pp. 2398-2401). International Speech Communication Association.

Towards Hierarchical Prosodic Prominence Generation in TTS Synthesis

Leonardo Badino[1,2]*, Robert A. J. Clark*[2]*, Mirjam Wester*[2]

[1] RBCS, Istituto Italiano di Tecnologia, Genova, Italy
[2] CSTR, The University of Edinburgh, UK

leonardo.badino@iit.it, robert@cstr.ed.ac.uk,mwester@inf.ed.ac.uk

Abstract

We address the problem of identification (from text) and generation of pitch accents in HMM-based English TTS synthesis. We show, through a large scale perceptual test, that a large improvement of the binary discrimination between pitch accented and non-accented words has no effect on the quality of the speech generated by the system. On the other side adding a third accent type that emphatically marks words that convey "contrastive" focus (automatically identified from text) produces beneficial effects on the synthesized speech. These results support the accounts on prosodic prominence that consider the prosodic patterns of utterances as hierarchical structured and point out the limits of a flattening of such structure resulting from a simple accent/non-accent distinction.

Index Terms: speech synthesis, HMM, pitch accents, focus detection

1. Introduction

This paper addresses the problem of identification of natural patterns of prosodic prominence and their generation in HMM-based English Text-to-Speech (TTS) synthesis.

It is commonly held within the speech synthesis community that the synthetic generation of prosodic prominence (and more in general of prosody) fails to sound appropriate in long utterances and/or when utterances are in context, making TTS synthesis not yet satisfactory in some applications like book reading and automated spoken dialogues.

Two are the main causes of such failure: i) in state-of-the-art TTS systems prosodic prominence patterns are predicted from text by relying on features that very loosely account for the combined effects of syntax, semantics, word informativeness and salience (given by the context) on prosodic prominence; ii) the prediction of prosodic prominence patterns is almost always approximated to the prediction of sequences of accented and non-accented words (or more in general of prominent and non-prominent words).

According to one of the the most popular theory of prosody, the Autosegmental-Metrical theory ([1], [2] and [3]), every utterance has a stress pattern (where the sentence-level stress must not be confused with lexical stress) that "reflects a set of prominence relations between the elements of the utterance" ([4]). Thus prosodic prominence is assumed to be a relative property of speech. The stress pattern is organized in a binary tree structure where two siblings are always tied by a weak-strong relation that states which of the two is most prominent. Figure 1 shows an example of prominence tree for the noun phrase "labor union" where the first syllable of "labor" is stronger (i.e., perceived as more prominent) than the first syllable of "union".

This hierarchical representation assumes that several degrees of prominence can be perceived.

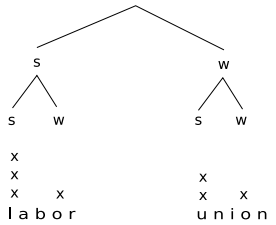

Figure 1: *Prominence tree of "labor union". The letter 's' and 'w' stand for relatively strong and relatively weak sibling respectively. The prominence tree can be mapped into a prominence grid in which the number of x's indicates the level of prominence of a syllable.*

The prominent syllables (i.e., the "strong" siblings, at any level of the tree) can be marked by a pitch accent while "weak" syllables (i.e., the "weak" syllables at the first level of the tree) can not. The accent perceived as most prominent accent of the utterance is usually referred to as "nuclear accent". It usually marks the focused word/phrase of the sentence, i.e., the most salient item (e.g., in the question-answer pair: "Who did that?","Paul did that", Paul is the focused word).

In this paper we show the limits that a simple accent/non-accent distinction implies in HMM-based TTS synthesis and the benefits coming from using a three-level accentuation. The third accent type is a "contrastive" nuclear accent, i.e., an accent that marks items that contrast with other words/phrases explicitly given by the discourse context (as in, e.g., "John wanted to talk to **Kate**, but had to talk to **Paul** first."). Note that in some theoretic accounts on focus (see [5]), focus and contrast are the same thing. Here we use a more intuitive and specific definition of contrast. This type of contrast is usually marked by a particularly strong accent, which has been claimed to have phonetic peculiarities. Either these phonetic peculiarities exist or not the contrastive accent is always at the top of the prominence tree.

There is very little work attempting to generate contrastive accents and more in general contextually appropriate prosodic prominence. Perhaps the closest previous study is [6] in which (manually annotated) contrastive accents were modelled with ToBI labels. The ToBI labels were used to predict F0 and duration which in turn were used as specification features in the cost function of a unit selection TTS system.

Part of the work presented here continues the work described in [7], with respect to which we show a largely improved generation of contrastive accents. Additionally in this paper we add evidence that a large improvement of the binary discrimination between pitch accented and non-accented words has no effect on the quality of the speech generated by the system.

2. Prediction of prominence patterns from text

To predict prosodic prominence patterns from text we built two components: a standard pitch accent predictor, and a contrastive accent predictor.

2.1. Pitch accent prediction

We trained and tested several accent predictors on the f2b subset of the Boston University Radio News corpus. A detailed description of these predictors is given in [8].

The set of training features extracted consists of a set of the most predictive features from the literature (which, for sake of simplicity we call "old" set) and a set of novel features ("new" set). The old features are: (i) Information Content (IC) = $-log(p(w_i))$, where $p(w_i)$ is the probability (computed offline using a 9-million words text corpus) of the word in position i;(ii) Relative IC (RIC) = $-log(p(w_i|w_{i-1}))$; (iii) Inverse RIC = $-log(p(w_i|w_{i+1}))$; (iv) Part-of-Speech (POS) of w_i; (v) length of w_i in number of characters.

The new proposed features are the following:

- Cached Information Content (CIC). CIC is a dynamic version of IC where $p(w_i)$ changes depending on whether w_i previously occurred in the text. CIC was computed by using the Cache Language Model functionality of the SRILM ([9]). For details on Cache Language Models see [10]. CIC takes into account the giveness of a word which has been claimed to be related to pitch accenting.
- Information Content of a Concept (ICC): this feature was proposed by [11] to measure the semantic similarity of the concepts associated to words. The lower the ICC of word w_i, the more specific the meaning of the concept associated to it. ICC accounts for the fact that specific words are more prone to accentuation than generic words (e.g., "cat" vs. "animal").
- Syntactic Dependencies (SD): a dependency tree is automatically extracted (using MaltParser [12]) and then features are extracted from it. A dependency tree indicates syntactic pairwise relations between a *dependent* and a *head* as shown in figure 2. Examples of these features are: name of the dependency in which the current word is a dependent (e.g., *Subject-of*); path and length of path between w_i and w_{i+1}. Some of these features are intended to account for a possible mapping between the syntactic relation linking two consecutive words and their prosodic prominence relation (i.e., the edge in the prominence tree), which in turn might affect word accenting.

To build the pitch accent predictor we experimented with a number of machine learning techniques already used in previous work on accent prediction: Classification and Regression Tree (CART, [13]), Bagging using CARTs as basic learners ([14]), Hidden Markov Models (HMMs) (where CART is used to compute the observation probabilities as, for example, in [15]), and Conditional Random Fields (CRFs) [16]. The use of the first two classifiers implies that the accenting of w_i does not depend on the accenting of the previous words (see [17] for example), while the use of the last two does not imply such independence.

The Bagging of CART method (which is a committee of CARTs) gives the best results with a 85.2% accuracy in a 10-fold cross-validation. To our knowledge this predictor compares favourably with all accent predictors (based on textual features only) proposed in previous work. The new features produce a 0.9% relative accuracy increase.

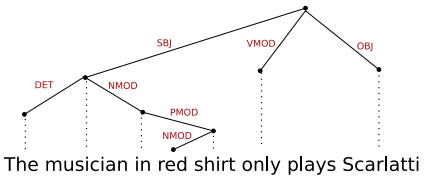

Figure 2: *Dependency tree of the sentence "The musician in red shirt only plays Scarlatti"*

2.2. Contrastive accent prediction

The contrastive accent predictor was trained on a set of excerpts of transcriptions of spontaneous dialogues in which there are pairs of words explicitly contrasting with each other that had been marked with particularly prominent accents. All these contrastive word pairs can be identified by looking at text only (e.g., "and, you know, even the **public** schools are behind the **parochial** schools").

The contrast predictor examines all the word pairs in the sentence that share the same gross POS (i.e., verb, noun, etc...). It is a Support Vector Machine-based binary classifier. The training feature set contains semantic, syntactic and morphological features. In a leave-one-out evaluation the classifier achieves a 74.4% precision and 22.4% recall in contrast identification.

Note that actually the classifier is very precise in recognizing word-pairs that are "semantically" contrastive (i.e., whose contrast can be recognized by looking at the text only) but some of these pairs had not been prosodically marked by the speakers and so the predicted classification is considered wrong. This problem could be largely reduced if the contrastive accent prediction task were carried out in two steps: first, identification of semantically contrastive words, and then prediction of contrastive accents (based on the semantic contrast identification). Unfortunately, at the moment, there are no corpora that would allow to train a semantic contrast tagger (or even better, a focus tagger). In [8] we propose active-learning- and semisupervised learning-based tools to effectively train contrast predictor on small corpora. More detailed description of the classifier and discussion can be found in [7] and in [8].

Some of the sentences containing contrastive word-pairs successfully identified in a leave-one-out validation by the predictor are used as testing sentences in the Experiment 2 described in this paper.

3. Prosodic prominence generation

Two large-scale experiments were carried out to: 1) assess the utility of an improved pitch accent predictor in a HMM-based TTS system that uses pitch accent as one of its linguistic features; 2) assess the naturalness of the generated contrastive accents and test whether introducing contrastive accents produces improvement in the overall speech quality. See [8] for more details.

Any modification to the original HMM based TTS system ([18]) for the generation of prosodic prominence was only ap-

plied at the level of the set of training linguistic features used by the system.

3.1. Experiment 1

We used two TTS systems that we named HTS05 and HTS05-PP. System HTS05 is the HMM based TTS system described in [18]. It uses the pitch accent predictor integrated in Festival whose accuracy on the BURN corpus is very poor (below 70%) as reported in [19]. System HTS05-PP uses our best pitch accent predictor, i.e., the Bagging-based predictor. Additionally HTS05-PP uses a large set of prosodic prominence related features that are not used by HTS05. This feature set includes:

- accentuation value ($\in$\{accent,no-accent\}) of \{previous,current,next\} phoneme. A phoneme is accented if it belongs to the nucleus of a stressed syllable of an accented word. Note that in HTS05 the accentuation value only refers to the syllable, i.e., it is only at the syllable level.
- uncertainty of the pitch accent predictor (only at the word level).
- features used by our pitch accent predictor (only at the word level)

There are at least two reasons to use the training features of the pitch accent predictor. First, these features are speaker-independent while the pitch accent predictor is not. Having speaker-independent features related to pitch accenting may be useful as our pitch accent predictor and HTS05-PP were trained on two different voices. Second, some of the training features (e.g. SD) of the pitch accent predictor may capture weak-strong prominence relations between two adjacent words that cannot be captured using the accent feature only.

Both systems were trained on 2025 utterances of the Roger voice from the Blizzard 2008 speech data set [20].

A first informal listening of a large number of synthesized utterances showed almost no difference between the two systems. This result was confirmed by two listening tests involving 30 British English native speakers. The first test was a preference test, where the subjects had to indicate their preference (or no preference), in terms of overall speech quality, for a HTS05 or a HTS05-PP generated utterance. We selected the 12 utterance pairs where the two systems sounded (to us) most different and presented them twice (in both orders), in random order, to each subject. The second test is a similarity test where the subject had to indicate which of the two synthesized utterances sounded more similar to the natural utterance.

Both tests showed no significant difference between the two systems, and even a preference, although not significant, was found for system HTS05 in the preference test (227 vs. 212, and 281 "no-preferences"). Such results show that improved pitch accent prediction alone is not sufficient to improve the overall speech quality and naturalness of a TTS system.

A possible explanation for these results is that a prosodic prominence model that mainly relies on the simple distinction between accents and non-accents is a far over-simplistic model of prosodic prominence that "over-flattens" the complex hierarchical structure of prosodic prominence.

However we can not draw a definitive conclusion yet as only one training voice has been used and, additionally, that voice is different from that on which the pitch accent predictor was trained. Perhaps more definitive conclusions could be drawn if the pitch accent annotation of the training speech dataset of the TTS system were manual.

3.2. Experiment 2

In this experiment only one TTS system was used (system HTS05-PP-E). The system is quite similar to the system described in [7] (although, as we will see, it produces much better results). It was trained on a speech dataset consisting of the same 2025 "neutral" style utterances used to train systems HTS05 and HTS05-PP plus all the 1683 utterances of the "carrier sentences" corpus [21]. This corpus contains hundreds of contrastive words recorded in three different templates as in the following example:

S1: *It was JAMES who did it.*
S2 *No, it was JOHN who did it!*
S3 *It was JOHN, not JAMES.*

The templates were repeated tens of times using different proper names. The speaker was asked to emphasize the names so that the contrastive accents are actually not "spontaneous" contrastive accents and are often emphatic.

Compared to the training data used in [7] the training data used in this experiment contains about 900 "neutral" utterances (i.e., uttered in a neutral, non-emphatic style) more as we found out that adding more neutral data helped smoothing the emphasis of the contrastive accents.

Concerning the linguistic feature set, HTS05-PP-E used the linguistic features used in HTS05-PP plus a set of features used to generate contrastive accents:

- \{previous,current,next\} phoneme emphasis value
- \{previous,current,next\} syllable emphasis value
- emphasis dependent name of the \{previous,current,next\} phoneme (e.g., emphatic /a/)
- emphasis dependent name of the syllable nucleus

There were three possible emphasis values: 0 if the word was not emphatic, A if the word was the first or the only emphatic word in the utterance, and B if the word was the second emphatic word in the utterance. In [7] only two emphasis values were used: emphatic and non-emphatic. We increased the number of emphasis values since an informal analysis of the previous system showed that the two contrastive accents were perceived as too similar and a differentiation seemed preferable.

Two different perceptual tests were designed: a preference test and a contrastive word detection test. In the first test we used 20 sentences from the sentence test used to train and test the contrastive accent predictor. The 20 sentences were selected from the set of sentences in which the predictor correctly identified contrastive word pairs.

We then synthesized two different versions of the same sentence using HTS05-PP-E. In one version (version StdC), contrastive words were accented with a standard pitch accent while in the other version (version EmphC) the contrastive words were accented with a contrastive accent (either value A or B depending on the contrastive word). The test participants were asked to indicate which version sounded best (the "no-preference" option was also available).

In the contrastive word detection test we selected 10 sentences containing at least one contrastive word pair and synthesized them with a contrastive accent on only one word (that could be a word of the contrastive pair if the sentence contained more than one contrastive word pair). The remaining words were normally accented by the accent predictor. The subjects were asked to indicate the word they perceived as most prominent. The subjects of this experiment were the same as those of Experiment 1.

EmphC	StdC	No preference	p-value_1	p-value_2
221	180	199	$p < 0.05$	$p = 0.094$

Table 1: *EmphC vs. StdC. In EmphC the contrastive words are marked with an emphatic contrastive accent while in StdC the same contrastive words are marked with a standard pitch accent. The first two columns show the number of preferences for one of the two versions and the third columns the number of cases in which subjects expressed no preference. 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 into two halves and summing one half to the EmpC preferences and the other half to the StdC preferences.*

The presence of contrast (identifiable from text) in the test sentences had the aim of making the emphasis recognition task more difficult by giving to the listeners no textual cues or misleading textual cues about the placement of the emphatic accent.

The results of the detection test show that contrastive accents were often clearly identifiable. In fact in 6 out of 10 utterances the number of speakers able to identify the most prominent word was significantly (with $p \ll 0.01$ in a binomial two-sided test) greater than the chance level (where the chance level was computed taking into account the emphatic word and all the accented words in the utterance).

In the preference test, in addition to listening to the utterances the subjects had to read the dialogue excerpts containing the test sentences. Results (Table 1) show a significant preference for the EmphC version supporting the hypothesis of the necessity of more than two levels of accentuation.

Note that in order to generate contrastive accents we needed accent so phonetically strong to be perceived as contrastive when "inserted" in any "neutral" context independently of the prominence of the words in that context. This approach may contrast the assumption that prosodic prominence is a relative property of speech and that the prominence of a word (i.e., the number of *x*'s in Figure 1) does not necessarily have specific phonetic correlates and its phonetic appearance may only be given by the the phonetic context. Unfortunately handling such kind of relative properties is not feasible in a standard HMM-based system where the phonetic realization of a hidden state (i.e., of a phone segment) is independent of the realizations of the preceding hidden states.

4. Conclusions

In this paper we addressed the problem of identification of natural patterns of prosodic prominence and their generation in HMM-based English TTS synthesis. We showed results that cast doubts on the actual utility of accurate binary accent prediction for TTS synthesis. Such non-utility may point out the limits of an oversimplified prosodic prominence model which flattens the hierarchical prosodic prominence structure. We showed that a less flatten model, obtained by introducing contrastive accents, improves the overall quality of the synthesized speech.

5. Acknowledgements

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (http://www.ecdf.ed.ac.uk/).

6. References

[1] Liberman, M., "The intonational system of English", PhD thesis, MIT Linguistics, Cambridge, MA, USA, 1975.

[2] Bruce, G., "Swedish word accents in sentence perspective", Developing the Swedish intonation model, Working Papers, Department of Linguistics and Phonetics, University of Lund, 1977.

[3] Pierrehumbert, J., "The Phonology and Phonetics of English Intonation", PhD thesis, MIT, Cambridge, MA, USA, 1980.

[4] Ladd, D.R., "Intonational Phonology" Cambridge University Press, Cambridge, UK, 1996.

[5] Umbach, C., "On the notion of contrast in information structure and discourse structure", Journal of Semantics, 21, 1–21, 2004.

[6] Pitrelli, J. and Eide, E., "Expressive speech synthesis using American English ToBI: questions and contrastive emphasis" Proceedings of IEEE ASRU, St. Thomas, Virgin Islands, 2003.

[7] Badino, L., Andersson, J., Yamagishi, J., and Clark, R., "Identification of contrast and its emphatic realization in HMM-based speech synthesis", in Proc. of Interspeech, Brighton, U.K, 2009.

[8] Badino, L., "Identifying Prosodic Prominence Patterns for English Text-to-Speech Synthesis", PhD Thesis, University of Edinburgh, 2010.

[9] Weintraub, M. et al., "Lm 95 project report. fast training and portability" Language Modeling Summer Research Workshop Technical Report, Research Note 1, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 1995.

[10] Kuhn, R. and De Mori, R.,"A cache-based natural language model for speech recognition", IEEE Transactions on Pattern Analysis And Machine Intelligence, 12(6), 1990.

[11] Resnick, P., "Using information content to evaluate semantic similarity in a taxonomy", in Proceedings of IJCAI, Montreal, Canada, 1995.

[12] Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kbler, S., Marinov, S., and Marsi, E., "Maltparser: A language-independent system for data-driven dependency parsing", Natural Language Engineering, 13(2):95–135, 2007.

[13] Breiman, L., Friedman, J., Olshen, R., and Stone, P., "Classification and regression trees", Wadsworth International Group, Belmont, CA, USA, 1984.

[14] Breiman, L., "Bagging predictors", Machine Learning, 24(2):123–140, 1996.

[15] Sun, X. and Applebaum, T., "Intonational phrase break prediction using decision tree and n-gramm model", in Proc. of Eurospeech, Aalborg, Denmark, 2001.

[16] Lafferty, J., McCallum, A., and Pereira, F., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data" in Proc. of ICML, Williamstown, MA, USA, 2001.

[17] Yuan, J., Brenier, J., and Jurafsky, D., "Pitch accent prediction: Effects of genre and speaker", in Proc. Interspeech, Lisboa, Portugal, 2005.

[18] Zen, H. and Toda, T., "An overview of Nitech HMM-based speech synthesis system for Blizzard challenge", in "Proc. of Interspeech", Lisboa, Portugal, 2005.

[19] Sridhar, V., Bangalore, S. and Narayanan, S., "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework", IEEE Transactions on Audio, Speech, and Language processing, 16(4):797–811, 2008.

[20] Karaiskos, V., King, S., Clark, R., and Mayo, C.,"The Blizzard challenge 2008", In Proc. Blizzard Challenge Workshop, Brisbane, Australia, 2008.

[21] Strom, V., Nenkova, A., Clark, R., Vasquez-Alvarez, Y., Brenier, J., King, S., and Jurafsky, D., "Modelling prominence and emphasis improves unit-selection synthesis", in Proc. Interspeech, Antwerp, Belgium, 2007.