Modelling Prominence and Emphasis Improves Unit-Selection Synthesis

Volker Strom[1], *Ani Nenkova*[2], *Robert Clark*[1], *Yolanda Vazquez-Alvarez*[1], *Jason Brenier*[2], *Simon King*[1], *Dan Jurafsky*[2]

[1]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[2]Linguistics Department, Stanford University, Stanford, CA

Abstract

We describe the results of large scale perception experiments showing improvements in synthesising two distinct kinds of prominence: standard pitch-accent and strong emphatic accents. Previously prominence assignment has been mainly evaluated by computing accuracy on a prominence-labelled test set. By contrast we integrated an automatic pitch-accent classifier into the unit selection target cost and showed that listeners preferred these synthesised sentences. We also describe an improved recording script for collecting *emphatic* accents, and show that generating emphatic accents leads to further improvements in the fiction genre over incorporating pitch accent only. Finally, we show differences in the effects of prominence between child-directed speech and news and fiction genres.

Index Terms: speech synthesis, prosody, prominence, pitch accent, unit selection

1. Introduction

Better synthesis of natural prosody is a crucial prerequisite for the development of more flexible and more natural text-to-speech systems. A key task in prosodic naturalness is assignment of *prominence*, such as the location of a *pitch accent* on a word or syllable [1, 2, 3, 4, 5].

Most previously published studies of prominence prediction have evaluated their algorithms in a very indirect manner, by seeing how often the model predicts the same prominence level on a word in a spoken sentence as a human labelling of the sentence. For example, while many (internally reported) tests were run at AT&T by the first author and others, no published results seem to have confirmed the hypothesis that advanced prominence prediction algorithms lead to perceptually significant improvements in synthesis quality. Indeed, many large-scale working systems rely on very simple prosodic algorithms, such as merely distinguishing (units in) function words from (units in) content words.

As a consequence, while the prominence prediction literature reports sophisticated algorithms which match the pitch accent labels in human-annotated corpora better than just distinguishing function and content words, we do not really know if these algorithms will improve the quality of synthesis. In addition, without embedded evaluation we cannot know what *kind* of prominence would help in synthesis. Standard pitch accents occur in every intonational phrase, but it has been suggested that particularly strong ('emphatic') accents might be helpful in synthesising contrastive accent (focus), as well as some aspects of child-directed speech (such as in storybook contexts). Is it therefore better to use an emphatic accent predictor in addition to a pitch accent predictor? If emphatic accents are indeed useful to model, how should we record the units in our unit database, to enable synthesis of emphatically accented words?

In order to address these questions, we recorded a new unit selection voice using scripts that had a variety of emphatic words (marked for the voice talent with capital letters). We then ran an automatic pitch accent predictor on the unit database. We synthesised a variety of test utterances in which pitch accents were assigned by the same automatic pitch-accent predictor. Emphatic accents were placed on capitalised words (as given by the authors in works of fiction), as well as on sentences from child-directed speech where the emphasis was used by the speaker in actual production. We controlled all other factors, systematically manipulating only the presence of the prominence models. In summary our experiments were designed to answer four questions:

1. Does fully automatic labelling of pitch accent (both in the unit database and the test sentences) lead to improved synthesis compared to the default voice with a baseline (function word/content word) prosodic model?
2. Is our method for recording contrastive sentences and encouraging the voice talent to produce emphatic prominence sufficient for adequate synthesis of emphasis?
3. Are pitch accent and emphatic accent complementary, improving different aspects of synthesis?
4. Do the prominence models benefit some genres more than others?

2. Corpus description

Our initial unit selection database was the seven hour speech corpus described in [6], which is comprised of Lewis Carroll's children's stories, a word list, newspaper and specialised texts. For the experiments we report here, we augmented this corpus with more newspaper text, the Arctic corpus [7] and carrier sentences for emphatic words as follows:

Lewis Carroll: Emphatic words in the Lewis Carroll portion are in all capital letters, giving a natural labelling of the corpus (194 emphatic tokens across 89 word types)

Word list: A list of 2,880 words selected for diphone coverage in the phrase-final syllable, each read five times as:

Ace, ace, ace. Ace? Ace!

which covers continuation rise (L-H% at the commas), terminal intonation (L-L% at the period and exclamation mark) and interrogative intonation (H-H% at the question mark). The speaker was asked to emphasise the last word.

Carrier sentences: The recording of word lists as described above still resulted in many missing diphones [6]. The reason for this is that emphasis is realised mainly on the lexically stressed syllable, which in polysyllabic words is not necessarily the last one. Furthermore the emphasised words in the word lists were often much less prominent than emphasised words in the Lewis Carroll part of the recordings and thus not reliable enough to render emphatic accents. To address this shortcoming, we recorded 1,683 sentences using the following templates designed to elicit emphatic productions both on phrase breaks and inside an intonation phrase:

“It was ERWIN who did it!”
“No, it was ELIZA who did it!”
“It was ELIZA, not ERWIN!”

To fill in the slots for emphatic words, we used 1,122 names selected from a list of 10,000 most common first and last names such that all diphones occurring in stressed syllables are covered. The list was then divided into pairs of names, and each pair gave one instance of the template shown above.

2.1. The automatic pitch accent predictor

To label the corpus for pitch accent, we used the prominence predictor described in [8]. The predictor is based on a single lexicalised feature, accent ratio:

$$AccentRatio(w) = \begin{cases} \frac{k}{n} \texttt{ if } B(k,n,0.5) \leq 0.05 \\ 0.5 \texttt{ otherwise} \end{cases}$$

where k is the number of times word w appeared accented in the corpus, n is the total number of times the word w appeared, and $B(k,n,0.5)$ is the probability (under a binomial distribution) that there are k successes in n trials if the probability of success and failure is equal. Accent ratio is the estimated probability of the word being accented, if that is significantly different from 0.5, and equal to 0.5 otherwise. This performs well on both spontaneous speech and news, outperforming prominence prediction algorithms based on part of speech, n-gram features, or even hand-labelled information status features such as given/new [8].

The accent ratio dictionary was compiled from two corpora annotated for pitch accent: Switchboard [9] and the Boston University Radio News corpus [10] and contains all words that have probability of being accented significantly different from 0.5. Words with a probability of being accented lower than 0.38 were marked as bearing no accent, and all other words were marked as accented[1]. This rule was used to annotate the corpus, with no use of any acoustic information at all.

2.2. Target cost

In the Festival *Multisyn* engine [11], prosody is modelled on the symbolic level only, with no explicit specification of phone durations or pitch targets. Instead, the target cost function imposes a penalty if the prominence labels of target and candidate unit do not match. In our current baseline system, there is no emphasis or accent component; word prominence is modelled only through a target cost component that distinguishes between function and content words. Adding an extra component to the target cost has the side effect of reducing the relative weight of other target cost components. Therefore, control of prosody comes at the potential cost of lower segmental quality.

3. Test materials

Test sentences were selected from the genres *news*, *fiction* and *child-directed speech*. These vary in degree of expressiveness and prosodic variation. Example sentences are shown in Table 1. News is the least expressive genre, with no emphatic productions. For this reason, eight news sentences were used to test the contribution of pitch accent only compared to default synthesis.

news:	The police said a loaded gun belonging to the suspect was recovered.
news:	Three uncertainties cloud the outlook for the world economy at the turn of the year.
news:	They feel they can’t go forward in this trial ethically.
emma:	Even YOUR satisfaction I made sure of.
emma:	Her friends evidently thought this good enough for her; and it WAS good enough.
alice:	He CALLED it a helmet, though it certainly looked much more like a saucepan.
gurney:	It’s HARD to see because it’s GREEN, like the GRASS.
gurney:	Peripheral vision is a FACT we can put in our ARTICLE.
gurney:	That’s a SECOND fact we can put in our ARTICLE.

Table 1: Sample test sentences.

The fiction sentences come from Jane Austen’s *Emma* and Lewis Carroll’s *Alice’s Adventures in Wonderland* (sentences that were not part of the voice recordings). A total of 27 sentences containing at least one orthographically marked emphatic word were selected from these.

The child-directed test sentences come from the *Gurney* corpus of short educational stories for children read by one female speaker of American English. The stories are part of the voice of an interactive agent in an automatic tutoring system for children [12]. The speaker’s productions were manually labelled for emphatic accents, and 13 sentences with emphasis were selected for the test set. Child-directed speech is marked by more prosodic variation than adult-directed speech, with more exaggerated intonation contours and increased use of emphatic elements to signal novelty and importance.

4. Test design

The selected test sentences were synthesised in four different ways: with the default voice (*-emphasis-pitch*), with added target cost components for emphatic accent only (*+emphasis-pitch*), for pitch accent only (*-emphasis+pitch*), and for both pitch and emphatic accent (*+emphasis+pitch*).

It has been suggested to treat pitch accents as a three-class problem by adding an ”optionally accented” class for cases where either assignment would be acceptable. This might improve synthesis quality by removing unnecessary target cost penalties and allowing better sounding units to be chosen. We had previously run a pilot experiment to compare two different pitch accent predictors for the *-emphasis+pitch* configuration: the classic accent yes/no versus yes/optional/no. 7 out of 10 listeners had a clear preference, of which six chose the two-class predictor, so this was used for the main experiment.

Five pairwise combinations of prominence models were tested. In order to keep the listening experiments reasonably short, each subject was assigned 8 sentence pairs for each of the five conditions, using a random 40x40 Latin square. The ordering within pairs was randomised, balanced across subjects.

52 subjects (a mixture of British and American English speakers) were recruited for a 30 minute long perception ex-

[1] Thus words not in the accent ratio dictionary are marked prominent.

	-emphasis-pitch	-emphasis+pitch	p-value
news	186	230	p = 0.03488
fiction	91	191	$p < 0.0001$
gurney	52	82	p = 0.0184

Table 2: Subject preference for synthesis with pitch accent model, compared to the default prosody. Improvements are significant for all three genres (using a two-sided Binomial test).

periment conducted through a web browser in which they were presented with the written sentence and two audio files; emphatic target words were capitalised. Subjects were instructed to listen to the stimuli as often as they wanted (although once could suffice to make a decision) and in any order, then to make a forced decision of which version sounded more natural.

In addition, an emphasis recognition test was performed. 20 relatively short sentences were synthesised with *+emphasis-pitch*. Each subject was presented with 8 of these (in a balanced design) along with the written form. They were instructed to mark the single word they perceived as most prominent.

5. Results

5.1. Automatic pitch accent prediction improves synthesis

We compared our function-word/content-word baseline (*-emphasis-pitch*) against the automatic pitch accent predictions based on the accent ratio dictionary (*-emphasis+pitch*).

Subject preferences between these two voices are shown in Table 2. In all three genres (news, fiction and child-directed), the addition of the pitch accent model lead to a significant improvement. In news, which tends to be much less expressive than the other two genres, and where the baseline distinction between function and content words leads to reasonable predictions, the improvement is smaller than for the other texts but is still statistically significant. The fiction sentences benefit most from the addition of the pitch accent model.

We believe this result to be extremely important. Prior work on pitch accent prediction has focused on developing novel machine learning algorithms or sophisticated features for pitch accent assignment, but no published work reports on whether these new algorithms or features actually improve the quality of synthesis. Our experiments convincingly demonstrate that our pitch accent model can be successfully integrated into synthesis and outperforms the baseline.

5.2. Emphatic accent produced successfully

Emphatic accents are generally rare, occurring in 2% of utterances in conversational speech or in the three books we used for selecting test sentences. They serve as an embellishment, and while not very common, when they do occur, they change the semantics or more expressively realise contrasts in the sentence.

Tables 4a and 4b show subject preferences between the default voice (*-emphasis-pitch*) and voices with added emphatic accent only (*+emphasis-pitch*) and both pitch and emphatic accent (*+emphasis+pitch*). In both cases, there is a significant improvement over the default voice for the fiction genre, showing that even though emphatic accent is not that common overall, when it does occur and is realised properly, it can change the perception of the sentence. The positive results of preference for the voices with emphasis compared to those without suggest that the contrastive scenario script for the voice recording was sufficient for synthesis of emphasis.

89:9	When he DID^{2} speak again, it was in a deep growl^{2}.
43:10	And when you've^{2} once heard^{10} it you'll be QUITE^{9} content.
86:14	We had SUCH^{18} a thunderstorm last^{3} Tuesday.
81:10	The WHITE^{17} kitten^{1} had had nothing^{3} to do with it.
95:10	You should NOT^{20} go on licking your^{1} paw like that!
83:11	THAT^{19} is so hard^{4} that I fear I'm unable!
41:7	He CALLED^{9} it a helmet^{3}, though it certainly^{8} looked much more like^{2} a saucepan.
95:14	Compare Mr. Martin^{1} with either of THEM^{21}.
95:8	She is a woman that one may, that one MUST^{19} laugh at^{1}.
83:8	YOU^{15} confined to the society of the illiterate and vulgar all^{3} your life!
90:8	Especially when ONE^{19} of those two is such a fanciful^{1}, troublesome^{1} creature!
95:7	I am going this moment myself; and I think the sooner YOU^{18} go the better^{1}.
56:14	Even YOUR^{10} satisfaction I made sure^{7} of^{1}.
72:8	But as to my^{1} LETTING^{17} her marry Robert^{1} Martin^{2}, it is impossible^{1}.
100:14	Not that she WANTED^{20} him to marry.
86:13	THIS^{18} article won't be so^{3} hard to write.
85:20	That^{1} shouldn't^{1} be TOO^{17} hard^{1}.

Table 3: Word superscripts indicate the number of listeners who perceived that word as most prominent. First column: Ratio of recognition rate (%) to chance level (%); average ratio is 8:1.

Results of the emphasis recognition test independently confirm this finding: Table 3 shows that on average, the recognition rate is 8 times above the chance level. In our previous work before recording the additional contrastive scenario sentences, the recognition rate was only 2.8 times above the chance level [6].

A closer look reveals that the positive contribution of emphasis is expressed exclusively in the test sentences comprising the fiction genre. In the sentences from Gurney, the child-directed educational stories, there is no significant difference between productions with emphasis and those without. We discuss this difference in greater detail in the next section.

5.3. Pitch accent and emphasis benefits are cumulative

Based on the results we have discussed so far, we were able to conclude that synthesis with models of pitch and emphatic accents, both individual and joint, are preferable to a default voice with no explicit prominence model. We also want to know if the benefits of the two prominence models are cumulative, that is, if the model for emphatic accent improves significantly over the pitch accent model used in isolation. The subject preferences in Table 4c between a voice with pitch accent (*-emphasis+pitch*) only compared to a combined emphasis and pitch accent model (*+emphasis+pitch*) indicate that for all the combined test sentences, we cannot draw such a conclusion. Within genre though, there are significant preferences in different directions. In the fiction sentences, the combined model is preferred significantly more often than the model with pitch but no emphatic accent. In contrast, in the child-directed speech, the voice without emphasis is preferred more often, and the addition of emphasis causes the quality of productions deteriorate.

This profound difference between the two genres reappears again in Table 4d, which was designed to answer the question "If we could add only one of the pitch accent (*-emphasis+pitch*) or the emphasis (*+emphasis-pitch*) models, which one would be more beneficial?" In the fiction portion of the test sentences, incorporating emphasis leads to significant improvements com-

4a	-emphasis-pitch	+emphasis-pitch	p-value
combined	151	265	$p < 0.0001$
fiction	79	204	$p < 0.0001$
gurney	72	61	$p = 0.3860$

4b	-emphasis-pitch	+emphasis+pitch	p-value
combined	132	284	$p < 0.0001$
fiction	62	217	$p < 0.0001$
gurney	70	67	$p = 0.8644$

4c	-emphasis+pitch	+emphasis+pitch	p-value
combined	202	214	$p = 0.5897$
fiction	111	173	$p = 0.0003$
gurney	*91*	*41*	$p < 0.0001$

4d	-emphasis+pitch	+emphasis-pitch	p-value
combined	211	205	$p = 0.8064$
fiction	115	161	$p = 0.0067$
gurney	96	44	$p < 0.0001$

Table 4: Preferences for different combinations of prominence models. P-values are from two-sided Binomial test.

pared to adding pitch accent alone. The opposite is true for the child-directed Gurney sentences; adding a pitch accent model is preferable to adding emphasis capabilities.

5.4. Emphasis in child-directed speech is different from that in fiction

Previous studies have shown that prediction of pitch accent is robust across speakers and genres [13]. Features useful for one genre were shown to also be helpful for others in a comparison between conversational speech, read news and child-directed speech. Our results in Table 2 support these findings, showing significant improvements in quality as a result of the addition of a fully automatic pitch accent model across all three genres.

The findings on emphatic accent reported in Table 4, on the other hand, show a marked difference between the fiction and child-directed genres. In the child-directed scenario, the emphatic accent is problematic, leading to inferior renditions of the sentences. This difference is important and deserves further detailed investigation. One possible explanation of the result is that the recordings of contrastive sentences used for synthesis of emphatic accent are not adequate for emphatic child-directed productions. An alternative explanation is that the synthesis of emphatic child-directed speech was indeed successful, but the adult participants in the perception experiments dispreferred this type of speech. The participants were instructed to simply choose the rendition of a sentence that they thought sounded better, with no other explanations about the genres or the intended listeners. Preferences might be influenced by the fact that the emphasis in the fiction and child-directed speech serves very different purposes, in one case changing the semantics or focus of the sentences, and in the other more generally signalling new or important information and ends of intonational phrases.

6. Conclusions

We have demonstrated using human listening experiments that integrating a fully automatic pitch-accent prediction algorithm based on a single feature, *accent ratio*, into the target cost results in better unit selection synthesis. For more emphatic accents, we show a new method for including them in the recording script that gives further improvements in the recorded sentences. Finally, our results suggest that the emphatic prosody in child-directed speech may be qualitatively different from the prosody in adult emphatic speech and requiring a different synthesis approach.

7. Acknowledgements

We would like to thank Scottish Enterprise for generous support for this Edinburgh-Stanford Link project, and Roger Burroughes for his voice. Simon King holds an EPSRC Advanced Research Fellowship.

8. References

[1] J. Hirschberg, “Pitch Accent in Context: Predicting Intonational Prominence from Text,” *Artificial Intelligence*, vol. 63, no. 1-2, pp. 305–340, 1993.

[2] K. Ross and M. Ostendorf, “Prediction of abstract prosodic labels for speech synthesis,” *Computer Speech & Language*, vol. 10, no. 3, pp. 155–185, 1996.

[3] S. Pan and J. Hirschberg, “Modeling local context for pitch accent prediction,” in *ACL 2000*. Hong Kong: Proceedings of ACL’00, 2000, pp. 233–240.

[4] X. Sun, “Pitch accent prediction using ensemble machine learning,” in *Proceedings of ICSLP*, 2002.

[5] M. Gregory and Y. Altun, “Using conditional random fields to predict pitch accents in conversational speech,” in *ACL’04*, Barcelona, Spain, 2004.

[6] V. Strom, R. Clark, and S. King, “Expressive prosody for unit-selection speech synthesis,” in *Proc. Interspeech*, Pittsburgh, 2006.

[7] J. Kominek and A. Black, “CMU Arctic databases for speech synthesis,” in *Tech Report CMU-LTI-03-177*, CMU Pittsburgh, PA, 2003.

[8] A. Nenkova, J. Brenier, A. Kothari, S. Calhoun, L. Whitton, D. Beaver, and D. Jurafsky, “To memorize or to predict: Prominence labeling in conversational speech,” in *Proceedings of NAACL-HLT*, Rochester, NY, 2007.

[9] J. Godfrey, E. Holliman, and J. McDaniel, “SWITCHBOARD: Telephone speech corpus for research and development,” in *Proc. ICASSP*, 1992, pp. 517–520.

[10] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, “The Boston Unversity radio news corpus,” Boston University, Tech. Rep. ECE-95-001, 1994.

[11] R. A. Clark, K. Richmond, and S. King, “Multisyn voices from ARCTIC data for the Blizzard challenge,” in *Proc. Interspeech 2005*, Lisbon, Portugal, Sept. 2005.

[12] R. Cole, S. Pellom, B. Hacioglu, K. Movellan, J. Schwartz, S. Wade-Stein, D. Ward, and W. Yan, “Perceptive animated interfaces: First steps toward a new paradigm for human-computer interaction,” *Proc. of IEEE: Special Issue on Human Computer Multi-Modal Interfaces*, vol. 91, no. 9, pp. 1391–1405, 2003.

[13] J. Yuan, J. Brenier, and D. Jurafsky, “Pitch accent prediction: Effects of genre and speaker,” in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005.