Pitch Accent Prediction: Effects of Genre and Speaker

Jiahong Yuan[1], Jason M. Brenier[2], Dan Jurafsky[1]

[1]Stanford University, [2]University of Colorado at Boulder
jy55@stanford.edu, jbrenier@colorado.edu, jurafsky@stanford.edu

10.21437/Interspeech.2005-504

Abstract

To build a robust pitch accent prediction system, we need to understand the effects of speech genre and speaker variation. This paper reports our studies on genre and speaker variation in pitch accent placement and their effects on automatic pitch accent prediction. We find some interesting accentuation pattern differences that can be attributed to speech genre, and a set of textual features that are robust to genre in accent prediction. We also find that although there is significant variation among speakers in pitch accent placement, speaker dependent models are not needed in accent prediction. Finally, we show that after taking speaker variation into account, there is little room to improve for state-of-the-art classifiers on read news speech.

1. Introduction

Speakers of English produce certain words in an utterance with special intonational prominence. These *pitch-accented* words typically are realized with increased duration, intensity, and fundamental frequency. The prediction of pitch accent from text is important for achieving naturalness in TTS and interpreting discourse structure in ASU applications.

Many features are useful in accent prediction, including part-of-speech [1], word N-gram probability [2], the number of syntactic phrases that a word initiates or terminates [3], overall sentence position [4], word informativeness measures [5], and collocation with neighboring words [6].

In this study, we explore two additional factors that influence the way in which speakers assign pitch accents to words: *speech genre*, and *speaker variation*. We focus on differences in pitch accent placement among four prosodically labeled corpora and their effects on building robust pitch accent predictors across genres. Previous studies have shown linguistic variation to be significant across language styles in multiple levels. For example, Hirschberg (2000) found that read speech differs from spontaneous speech with regard to speech rate, intonational contour, disfluency, and prosodic correlates of discourse structure [7].

This paper is organized as follows: in the next section, we present the data used in our analyses and experiments. In section 3, we report the results of our analyses of the effect of genre on the distribution of pitch accent. In section 4, we discuss which features are robust to genre in accent prediction. In section 5, we examine speaker variation and its effects on accent prediction. Finally, in section 6 we present our conclusions.

Thanks to OSU for generously making available a pre-release of the Buckeye corpus. Thanks to Kristina Toutanova for helping on using the Stanford POS tagger. This research was partly funded by the Edinburgh-Stanford LINK and the NSF via IIS-0325399.

2. Corpora

Our research makes use of four independent prosodically labeled corpora representing two distinct speech genres: read and spontaneous.

Boston University Radio News Corpus [8] was used to represent the read news broadcast genre. A portion of the corpus was labeled using the ToBI transcription conventions, including all of the news stories (11203 words) from one female speaker as well as part of the news stories (1662 words) from five other speakers.

The other read speech corpus, *Gurney*, was developed from child-directed stories that were read aloud by a female native speaker. This speech serves as the voice of the interactive agent in an automatic tutoring system [9]. The corpus contains a total of 4 stories (1906 words) annotated with the ToBI intonation conventions.

A subset of the *Switchboard* corpus of conversational telephone speech [10] that was prosodically labeled using the Tilt Intonational Model [11] was used as one of our spontaneous speech corpora. It contains a total of 4762 words from multiple female speakers (the male data were excluded from our studies).

The last corpus is *Buckeye* [12], which consists of speech elicited from 40 middle-class, Caucasian natives of central Ohio in a spontaneous interview format. The resulting interviews are largely monologic speech from the interviewee. One interview (7148 words) from a female speaker was annotated using the ToBI prosodic conventions and used in our studies.

For each word in these corpora, we used the ToBI and Tilt labels to assign a binary value indicating whether the word was accented or not, and extracted various other features such as part of speech.

3. Genre Variation in Accent Placement

We analyzed the differences in accent placement among the above four corpora (female only) and their respective speech genres. Results reported below are statistically significant at the 0.01 level or better.

Table 1 lists the percentage of the words that are accented under each part-of-speech and in each corpus.

We see that function words are less likely to be accented than content words, confirming earlier studies (e.g. [1], [13]). This effect holds for each of the four individual corpora.

However, content words in the read speech style (*Boston Radio* and *Gurney*) are more likely to be accented than content words in the spontaneous speech style (*Switchboard* and *Buckeye*). Contrarily, function words in the read speech style are more likely to be unaccented than function words in the spontaneous speech style. This suggests that in read speech, broad part of speech category is a better predictor of accentuation than it is in spontaneous speech, where the

clear distinction between accented content words and unaccented function words is blurred. We can illustrate this point by introducing the concept of **accent ratio**, defined as the number of accented tokens of a word divided by the total number of tokens of that word in the corpus. As illustrated in figure 1, there are more words in spontaneous speech whose accent ratio is in the middle between 0 and 1; by contrast, words in read speech tend to have accent ratios closer to 0 or 1.

From Table 1, strong differences in average accent ratios within the adverbial (R), the determiner (DT), the coordinating (CC), and the exclamative (UH) word subclasses can be seen between read and spontaneous speech. Further analyses show that these differences result from both accent ratio and relative frequency differences for particular words in the word subclasses, as shown in Table 2. For example, the word *never* is always accented in read speech, but only 53% and 25% of this word's tokens are accented in the two spontaneous speech corpora, *Buckeye* and *Switchboard*, respectively. The definite article *the* is never accented in read speech and is infrequently accented in spontaneous speech. Although the accent ratio for this word does not vary greatly between the read and spontaneous styles, the minor difference that exists is magnified by the fact that *the* accounts for a proportion of all determiners in read speech that is twice that of spontaneous speech (57.1% and 57.9% in read speech vs. 21.4% and 33.2% in spontaneous speech). Filled pauses *uh* and *um* didn't appear in read speech but account for about 30% of all exclamatives in spontaneous speech. Interestingly, *um* is more likely to be accented than *uh*. (72% and 18% for *um* and 12% and 13% for *uh*). This result supports the proposal that speakers use *uh* and *um* to announce that they are initiating what they expect to be a minor (*uh*), or major (*um*), delay in speaking [14].

Table 1: Percentage of the words accented under each part-of-speech category and in each corpus

POS category		Boston	Gurney	Buckeye	SWBD
Content words	CD	.77 (316)	.64 (14)	.76 (63)	.78 (54)
	J	.84 (877)	.92 (157)	.81 (269)	.84 (253)
	N	.80 (3948)	.84 (825)	.81 (783)	.79 (647)
	R	.79 (416)	.84 (164)	.61 (583)	.57 (425)
	V	.59 (1805)	.56 (592)	.57 (1432)	.49 (844)
	Total	.75 (7362)	.75 (1752)	.66 (3130)	.64 (2223)
Function words	CC	.01 (366)	.13 (78)	.16 (353)	.18 (253)
	DT	.11 (1036)	.09 (302)	.25 (470)	.21 (433)
	IN/TO	.07 (1538)	.12 (327)	.16 (762)	.16 (490)
	MD	.24 (173)	.17 (58)	.31 (145)	.28 (43)
	P	.24 (483)	.25 (292)	.35 (1504)	.28 (715)
	UH	.00 (1)	1.00 (12)	.85 (552)	.56 (467)
	Other	.30 (244)	.48 (77)	.35 (232)	.40 (138)
	Total	.12 (3841)	.18 (1146)	.35 (4018)	.29 (2539)
Total		.54 (11203)	.52 (2898)	.49 (7148)	.46 (4762)

*The numbers in the brackets are the frequencies of the words under each category.

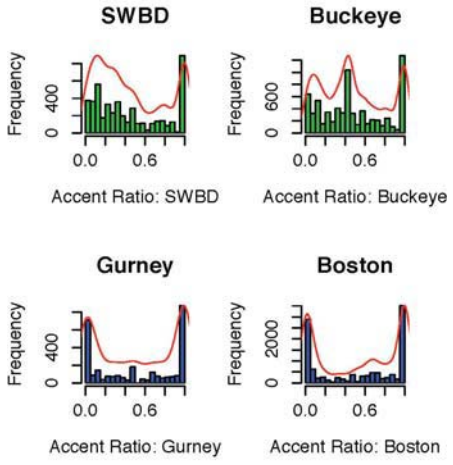

Figure 1:Distribution of pitch accent ratio (the number of accented tokens of a word divided by the total number of tokens of that word in the corpus)

Table 2: Pitch accent ratio and relative frequency of word in part-of-speech subclass (shown in the brackets below the accent ratio) of example words under each corpus

word	Boston	Gurney	Buckeye	SWBD
just (R)	.57 (1.6%)	.80 (3.0%)	.44 (15.2%)	.31 (6.8%)
never (R)	1.0 (.9%)	1.0 (1.2%)	.53 (2.9%)	.25 (.9%)
the (DT)	.00 (57.1%)	.00 (57.9%)	.01 (21.4%)	.02 (33.2%)
that (DT)	.42 (1.8%)	.44 (2.9%)	.38 (23.1%)	.42 (16.6%)
and (CC)	.01 (64.4%)	.12 (73.0%)	.12 (67.9%)	.15 (71.5%)
but (CC)	.00 (27.8%)	.06 (23.0%)	.31 (24.0%)	.23 (16.9%)
um (UH)	- (0.0%)	- (0.0%)	.72 (15.0%)	.18 (14.1%)
uh (UH)	- (0.0%)	- (0.0%)	.12 (4.7%)	.13 (23.9%)

4. Robustness to Genre in Accent Prediction

Although many features have been found useful in pitch accent prediction [1, 2, 3, 4, 5, 6], to our knowledge it has not been studied whether the features are robust across genres.

Using feature selection techniques we can find the best feature set of each speech genre in accent prediction. The robustness of a genre-independent common set of features can be calculated by comparing the performance of the common feature set with the performance of the best feature set of each genre. Following this strategy, we did feature selection experiments, add-one-in and leave-one-out, on each of the four corpora. The experiments were done using the decision tree package C4.5 [15] and 10-fold cross validation within each corpus, starting with the following features: F1: *part of speech*; F2: *unigram probability*; F3: *bigram probability*; F4: *backward bigram probability*; F5: the *position* of the word in an IP; F6: the *Information Content* (IC) of the word (calculated as the negative log likelihood of the word in a corpus); F7: the *accent ratio* of the word (calculated as the number of accented tokens of a word divided by the total number of tokens of that word in a corpus). Table 3 lists the results.

Table 3: The best feature sets and their classification error rates, selected by add-one-in and leave-one-out

corpus	add-one-in	leave-one-out
Boston	F7 (16.4%)	F1+F2+F4 (16.1%)
Gurney	F4+F5+F6+F7 (17.4%)	F4+F5+F6+F7 (17.4%)
Buckeye	F1+F2+F5+F6 (24.9%)	F1+F2+F5+F6 (24.9%)
SWBD	F1+F2+F5+F6 (24.0%)	F1+F2+F5+F6 (24.0%)

Used in isolation, the *accent ratio* feature (F7) is the most helpful for *Boston radio news*. However, it cannot be combined with any other features to generate a better classification result under this genre. The non-compatibility of the *accent ratio* feature with the other features is also seen in *Buckeye* and *Switchboard*. For example, if we add the *accent ratio* feature to the best feature set of *Switchboard*, the classification error rate goes up to more than 30% from 24%.

Excluding *accent ratio*, the remaining features, F1 - F6, perform nearly as well as the best feature set for each corpus. As listed in Table 4, the error rate of the model using these six features is less than 1% higher than the error rate of the best model for each corpus. This suggests that features F1 through F6 are a robust feature set across genres.

Table 4: Error rate difference between the robust feature set and the best set of each corpus

corpus	F1 through F6	best set	diff.
Boston	16.6%	16.1%	0.5%
Gurney	18.3%	17.4%	0.9%
Buckeye	25.1%	24.9%	0.2%
SWBD	24.0%	24.0%	0.0%

5. Effects of Speaker Variation on Accent Prediction

5.1. Speaker variation

A portion of the Boston University News corpus (lab news, 1662 words) was prosodically labeled for six speakers. These data were used for our speaker variation analyses.

Table 5 lists the percentage of the words that were either accented or unaccented by at least four, five, or six speakers. These numbers measure the consistency among the speakers in pitch accent placement. We see that only 59% of the words were consistently accented or unaccented by all the six speakers, suggesting that there is significant variation among speakers.

Table 5: Percentage of the words that were consistent among speakers in pitch accent placement

consistent among:	percentage of the words
all the six speakers	59.1%
at least five speakers	81.2%
at least four speakers	94.0%

We can see from Table 5 that 94% of the words were consistent among at least four of the six speakers. This means that 6% of the words were 'arbitrary' with regard to accent placement: for each of these words, half of the six speakers accented them whereas the other half did not.

Although arbitrariness of pitch accent placement is a factor contributing to speaker variation, it cannot explain why only 59% of the words were consistent among all the six speakers and only 81% of the words were consistent among at least five speakers. Some other factors like speaker sex may play an important role. For example, we find that among the six speakers, the males placed more pitch accents on the words (55%) than the females (49%). Further studies are needed to explore this issue.

Our study of speaker variation in pitch accent placement leaves two questions open: 1. Do we need a speaker-dependent model in accent prediction? 2. How does speaker variation affect the evaluation of accent prediction? These questions are addressed in the following sections.

5.2. Do we need a speaker-dependent model?

In the *Boston University News* corpus, a larger data set (9541 words) from one of the six speakers was also prosodically labeled. Training on this data set, we built a decision tree classifier using C4.5. We then evaluated the classifier on the data prosodically labeled for all the six speakers (1662 words). The word accuracy rates are listed in Table 6.

Interestingly, although trained on f2b, the classifier has better performance on the speakers m1b (86.4%) and m3b (84.8%) than on f2b (82.1%). On the other hand, the worst accuracy rate (80.1%) is only 2% lower than the accuracy rate on f2b. These results suggest that, at least for the radio news speech style, speaker-dependent models cannot improve the performance of a pitch accent classifier and hence are not needed.

Table 6: Accuracy rates of the classifier trained on f2b and evaluated on different speakers

evaluated on:	accuracy rate
f2b	82.1%
f1a	80.1%
f3a	80.9%
m1b	86.4%
m2b	82.2%
m3b	84.8%

5.3. Reevaluation of accent prediction performance

From Table 6, we can see that the accuracy rate of our accent classifier ranges from 80% to 86% on different individual speakers. The state-of-the-art accuracy rate numbers (tested on one speaker, f2b) reported in the literature are also in this range [2, 4]. If we take into account speaker variation, however, the performance of our accent classifier is greatly improved.

In Table 7 the accuracy rates were calculated in the following way: the classification, presence or absence of a pitch accent, is correct if it is the same as at least three speakers, at least two speakers, or at least one speaker. We see that 97.5% of the words are correctly classified when we test whether a word can be accented by any of the six speakers.

Table 7: Accuracy rates calculated against at least one, two, and three speakers.

calculated against:	accuracy rate
at least three speakers	89.7%
at lease two speakers	94.0%
at least one speaker	97.5%

We also evaluated the performance of the classifier in another way: First, we assigned an accent value to each word (1 if it is accented and 0 if not). Then we calculated the mean (expected) accent value of each word for the six speakers, as well as the Root Mean Square (RMS) difference between the accent values of each speaker and the mean accent values. The RMS difference can be seen as a measurement of divergence of the speaker from the expected accent placement. Finally, we calculated the RMS difference between the results of our classifier and the mean accent values. The RMS varies from .239 to .291 among the speakers, having a range of .052. As expected, the classifier has a greater RMS (.314) than the speakers (otherwise the classifier would be as perfect as a normal speaker). Nonetheless, its RMS value is only .023 greater than one of the speakers (.291), much lower than the range of speaker variation (.052).

To summarize, when we take into account speaker variation, our classifier, while not perfect, has very little room to improve.

6. Conclusions

Content words in the read speech style are more likely to be accented than in the spontaneous speech style. Contrarily, function words in read speech are more likely to be unaccented than in spontaneous speech. These differences result from both accent placement and lexical choice.

There exists a set of features that are robust to genre in accent prediction. Some features like *accent ratio*, however, should be excluded from the set.

Although there is significant variation among speakers in accent placement, speaker dependent models cannot improve accent prediction and hence are not needed. Finally, after taking account of speaker variation, there is little room to improve for state-of-the-art classifiers on read news speech.

7. References

[1] Hirschberg, J., "Pitch accent in context: Predicting intonational prominence from text", *Artificial Intelligence*, 63:305-340, 1993.

[2] Ross, K. and Ostendorf, M., "Prediction of abstract prosodic labels for speech synthesis", *Computer Speech and Language*, 10(3): 155-185, 1996.

[3] Chen, K. and Hasegawa-Johnson, M., "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model", *Proc. of ICASSP*, 509-512, 2004.

[4] Sun, X., "Pitch accent prediction using ensemble machine learning", *Proc. of ICSLP*, 561-564, 2002.

[5] Pan, S. and McKeown, K., "Word informativeness and automatic pitch accent modeling", *Proc. of EMNLP/VLC*, 1999.

[6] Sproat, R., "English noun-phrase accent prediction for Text-to-Speech", *Computer Speech and Language*, 8:79-94, 1994.

[7] Hirschberg, J., "A corpus-based approach to the study of speaking style", In *Prosody: Theory and Experiment*, Horne, M., ed., 335-350, Kluwer Academic Publishers, Dordrecht, 2000.

[8] Ostendorf, M., Price P. J., and Shattuck-Hufnagel S., *The Boston University Radio News Corpus*, Boston University Technical Report ECS-95-001, Boston University, Boston, MA, 1995.

[9] Cole, R., van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W., and Yan, J., "Perceptive animated interfaces: First steps toward a new paradigm for human computer interaction", *Proceedings of the IEEE: Special Issue on Human Computer MultiModal Interface*, 91(9):1391-1405, 2003.

[10] Godfrey, J., Holliman, E., and McDaniel, J., "SWITCHBOARD: Telephone speech corpus for research and development", *Proc. of ICASSP*, 517-520, 1992.

[11] Taylor, P., "Analysis and Synthesis of Intonation Using the Tilt Model", *J. Acoust. Soc. Amer.*, 107:1697-1714, 2000.

[12] Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W., "The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability", *Speech Communication*, 45:89-95, 2005.

[13] Altenberg, B., *Prosodic Patterns in Spoken English: studies in the correlation between prosody and grammar for text-to-speech conversion*, Lund University Press, Lund, 1987.

[14] Clark, H. H., and Fox Tree, J. E., "Using uh and um in spontaneous speech", *Cognition*, 84:73-111, 2002.

[15] Quinlan, J. R., *C4.5: programs for machine learning*, Morgan Kaufmann Publishers, San Mateo, 1993.