Issues of Optionality in Pitch Accent Placement

Leonardo Badino, Robert A.J. Clark

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

Abstract

When comparing the prosodic realization of different English speakers reading the same text, a significant disagreement is usually found amongst the pitch accent patterns of the speakers. Assuming that such disagreement is due to a partial optionality of pitch accent placement, it has been recently proposed to evaluate pitch accent predictors by comparing them with multi-speaker reference data. In this paper we face the issue of pitch accent optionality at different levels. At first we propose a simple mathematical definition of intra-speaker optionality which allows us to introduce a function for evaluating pitch accent predictors which we show being more accurate and robust than those used in previous works. Subsequently we compare a pitch accent predictor trained on single speaker data with a predictor trained on multi-speaker data in order to point out the large overlapping between intra-speaker and inter-speaker optionality. Finally, we show our successful results in predicting intra-speaker optionality and we suggest how this achievement could be exploited to improve the performances of a unit selection text-to-speech synthesis (TTS) system.

1. Introduction

In this paper we propose a new evaluation function for evaluating pitch accent predictors and a novel approach that exploits the variability of pitch accent patterns in order to improve the prosodic realization of a unit selection TTS system. In natural speech, alternative prosodic realizations of a given utterance can be equally acceptable. Even when a speaker is required to utter a sentence in a specific standard speech style (that of radio news speakers, for example) she/he will be free to choose amongst different prosodic patterns without altering the meaning of the sentence [1]. This freedom of choice affects different aspects of prosody, ranging from prosodic phrasing to the intonation contour. This prosodic variability offers a further degree of freedom to the developers of speech synthesis systems (or at least to those using the unit selection technique) who want to create systems able to go beyond a neutral prosodic realization making them able to convey additional meaning through prosody. In unit selection, a predefined prosodic target is usually expressed by a sequence of symbolic values describing F0 and segmental duration. These prosodic values are included into the specifications of the target utterance. The target is matched by selecting the appropriate acoustic units and, in some cases, by applying signal processing techniques. In such a context, imposing one single predefined prosodic target can involve a large amount of speech processing and a drastic reduction of the unit search space, thus resulting in a poor quality speech production, usually less acceptable than that of a system not supported by any prosodic model. As a consequence, and taking into account the prosodic variability of natural speech, new "softer" approaches have been recently proposed, for example, in [2] alternative prosodic patterns are implemented into a weighted-finite-state-transducer (WFST), which is then composed with the WFST describing the segmental information of the acoustic database. The unit sequence with the best combined cost is chosen. Prosodic constraints can be further relaxed by dropping the idea of explicitly defining the allowed prosodic patterns and selecting an implicit prosodic model by relying on the inherent prosodic structure of the speech database [3]. In our work we focused on the variability of prosodic patterns looking at a single type of prosodic event: pitch accent. We first analyzed the section of the Boston University Radio News corpus [4] where speech data have been collected by recording different speakers reading the same sentences. We show, for any combination of speakers, the intra-speaker disagreement in placing pitch accents. Then, starting from previous work, we faced the problem of evaluating pitch accent predictors on multi-speaker data, assuming that the intra-speaker disagreement is mainly due to a high degree of optionality in placing pitch accents. Our solution implies a simple mathematical definition of optionality which led us to the formulation of a new evaluation function. Subsequently, we tested our main work hypothesis, that is the assumption that the optionality observed when comparing the prosodic realization of different speakers (intra-speaker optionality) largely overlaps with inter-speaker optionality, that is the optionality that would be found if a speaker repeatedly read the same text without changing is speaking style. We compared a pitch accent predictor trained on single speaker data with a predictor trained on multi-speaker data. From the high similarity of performances of both predictors we inferred the validity of our hypothesis. Finally, we found out that our definition of optionality was determinant in our successful attempt of predicting optionality and, supported by the high similarity of intra and inter speaker optionality, we devised a simple method to exploit this achievement in order to improve the prosodic realization of a unit selection TTS system that uses pitch accent prediction to model prosody.

2. Disagreement Among Speakers

A section of the Boston University Radio News (BURN) corpus contains the speech of six different speakers (3females: f1a, f2b, f3a, and 3 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 Figure1).This part of the BURN corpus was already analyzed in [5] to investigate the intra-speaker disagreement in pitch accent placement. However, here, we provide some further data, useful for our purposes. Figure 2 shows the percentages of intra-speaker agreement for each combination of speakers and the agreement mean, with respect to the

	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*

number of speakers involved, on a text of 1662 words. The vertical segments 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. These two percentages may suggest a correlation between degree of agreement and speaker genre, but if we look at all the 20 possible triplets of the six speakers we see that the combination with the highest agreement (77.61%) is f2b-m1b-m3b, which consists of one female and two males. We did not carry out any study to investigate which are the factors that correlate to intra-speaker agreement and to what extent, but from an informal analysis it seems that speaker profession (is she/he a professional speaker?) is at least as significant as speaker genre.When comparing the agreement among speakers in pitch accent placement we can compute the proportion of agreement that is not due to chance by using the Kappa statistics:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of times speakers agree and P(E) the proportion we would expect them to agree by chance. In our case, assuming that accent and non-accent are equiprobable (the percentage of accented words for this speech style ranges from 45% to 55%) the κ value for the six speakers is 0.57.

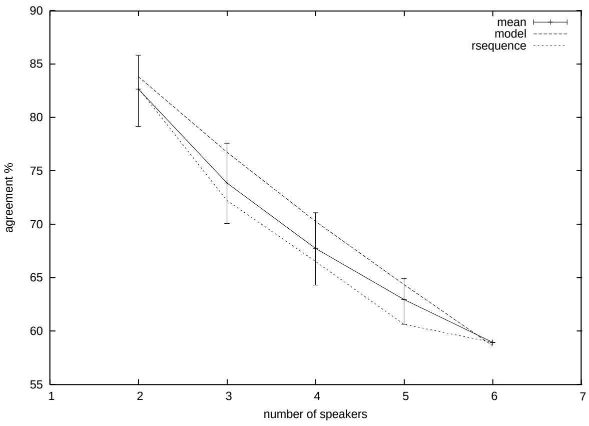

Figure 2: *Speakers agreement in pitch accent placement. The mean line represents the mean disagreement value. The rsequence line shows the disagreement resulting when adding a speaker in the order: f1a, f2b,f3a,m1b,m2b,m3b.*

3. Optionality and Pitch Accent Predictor Evaluation

3.1. Previous Works

If we make the assumption that when two or more speakers disagree in placing, or not, a pitch accent on a syllable, that pitch accent can be considered an optional accent, then we can reconsider the usual evaluation practice in which a pitch accent predictor is compared with only a single speaker. [5] and [6] used an evaluation function that considers a predicted event (accent or no-accent) wrong if it is not yielded by any of the speakers/annotators. Although the two works differ for the language (English vs Dutch) and the type of data test used (prosodically annotated speech vs prosodic labels directly derived from text) their conclusions are very similar: when optionality is taken into account in evaluating their automatic pitch accent predictors the performances of their predictors are very close to those of humans. This conclusion assumes that the optionality occurring when comparing speakers (intra-speaker optionality), is the same optionality that can occurs within a single speaker (inter-speaker optionality). As a consequence the accent pattern chosen by a speaker is made up of a compulsory part and an optional part, which can be exchanged with the optional part of (an)other speaker(s) without altering the coherence and naturalness of the whole accent pattern. There are however possible side-effects in this assumption. First, even if a pitch event is optional all the speakers can choose the same value. Second, the optional part of the pitch accent pattern of a single speaker can be related to the speaking style of the speaker herself/himself and, moreover, can be influenced by other factors that determine her/his speaking style, for example her/his speaking speed. As a consequence, mixing a speaker optional part with that of other speakers may result in an unnatural and "distorted" pattern. Finally, the evaluation function used in both works ignores a possible sintagmatic behavior of pitch accents: the placement of an accent can influence the placement of the following ones. In spite of that, in our work we kept the idea of evaluating accent predictors comparing them with multi-speaker data, supported by the fact that, as we will show later, fortunately, part of these side-effects is probably not so significant as it may seem at a first glance and can be reduced using a different evaluation function. Nevertheless, even assuming that these side-effects do not occur, the evaluation functions proposed in the previous works have still significant drawbacks. Figure 2 shows how the speaker agreement quickly decreases when the number of speakers increases. As a consequence it is easy to see how the evaluation function of [5] and [6] is strongly dependent on the number of speakers involved.

Figure 3 shows this fact by comparing three predictors (one of those is actually a speaker) varying the number of speakers involved in the test. The more the speakers in the test data are, the lower the intra-speaker agreement is and consequently the better the predictor results are. Consider the predictor A, which assigns a pitch accent to each words. If it is evaluated on six speakers, its accuracy rate is 73%, that means that we could build a predictor that accents the 73% of overall words, and performs a 100% of correct predictions. But, since the percentage of pitch accent in read speech ranges from 45% to 55% such a predictor is not appropriate to model pitch patterns of real speech. When looking at the speaker disagreement we should take into account that the steep decrease is partially due to the simple fact of adding new speakers even if the disagreement in each pair of speakers is low. In order to better illustrate that

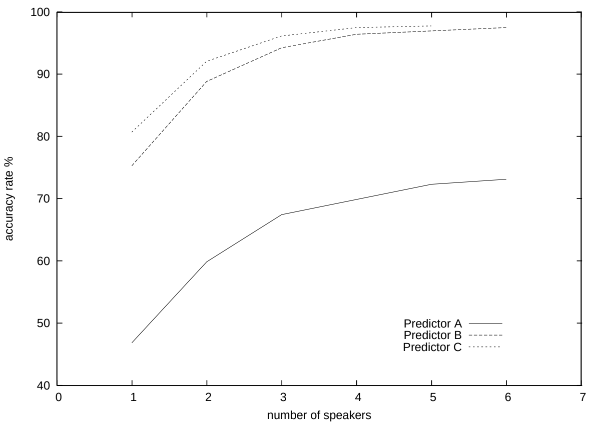

Figure 3: *Three predictors tested over different numbers of speakers. The sequence of speakers combination is f1a, f1a-f2b, f1a-f2b-f3a, f1a-f2b-f3a-m1b, f1a-f2b-f3a-m1b-m2b, f1a-f2b-f3a-m1b-m2b-m3b. Predictor A is an all-accented predictor. Predictor B is described in section 5. Predictor C is the speaker m3b.*

we could suppose that each word token in the test text has a non-zero probability of being optional, that is of being assigned both accent values (accented/non-accented) and that each pitch accent is independent from the others. If we assume p being the average probability of the most probable event for each word token, the agreement percentage can be modeled as:

$$(m1) \quad A(n) = 100[p^n + (1-p)^n]$$

where n is the number of speakers involved. In Figure 2 we plotted $A(n)$ (model) setting p to 0.9157. This value was obtained by imposing p^6 (the term $(1-p)^6$ was ignored) equal to the real agreement of six speakers (58.96%).

Even if our model is certainly approximate it clearly shows how even for high values of p the agreement percentage rapidly decreases by adding new speakers and gives a clue of what happens if more than six speakers are compared. Moreover this model allows us to see the intra-speaker optionality value not as a simple binary value but as a gradient one, which is a function of the probability of each word token of being assigned both pitch events. This concept is the base of our work.

The number of speakers is not the only parameter that can influence the predictors evaluation: the evaluation function of Figure 3 considers correct a pitch event if it is realized by at least one speaker, but we could be more strict and choose an evaluation function that marks as correct a predicted pitch event only if it is realized by more than one of the speakers involved. Considering n the number of speakers involved in the test and m (with $m < n$) the acceptable (for the evaluation function) number of speakers that realize the same pitch event of the predictor, we can write the evaluation function for each word token i:

$$OE(w_i) = \begin{cases} 1 & \textit{if at least m speakers realized} \\ & \textit{the predicted event} \\ 0 & \textit{otherwise} \end{cases} \qquad (1)$$

	$m = 1$	$m = 2$	$m = 3$
Predictor A	73.17	65.04	60.16
Predictor B	97.56	94.06	88.89

Table 1: *Accuracy rates of two predictors for different values of m ($n = 6$).*

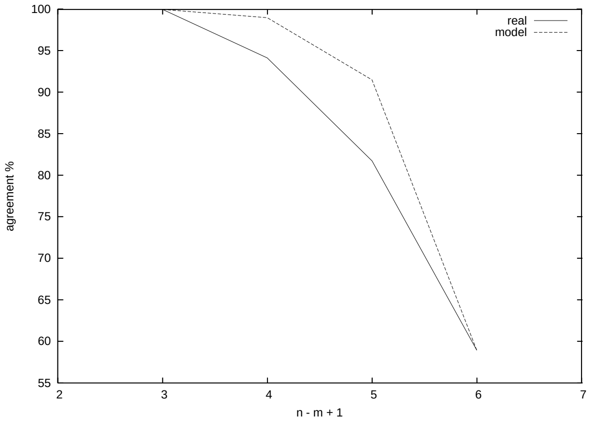

Figure 4: *Speakers agreement for different values of m (n=6)*

Table 1 shows the evaluation of two predictors already used in figure 3, this time always compared with all the six speakers ($n = 6$) but varying m. The high dependency of the evaluation function on m is again explained by the speaker disagreement: when m increases the number of cases in which the pretiction is considered correct independently on its value decreases. For example if $m = 1$ the prediction is always correct in all the cases where at least one speaker disagrees whereas it can be wrong or correct only when all the speakers agree. In figure 4 we plotted the percentage of pitch events that are consistent among all the six speakers (bottom right), at least five of the six speakers and so on. We also plotted an agreement function based on the same hypotheses made for (m1). Since the number of combinations of k speakers taken from a set of n speakers is given by $\left(\begin{array}{c} n \\ k \end{array}\right)$, in this case the agreement function is:

$$(m2) \quad A(n,m) = 100 \sum_{k=n-m-1}^{n} \left(\begin{array}{c} n \\ k \end{array}\right) [(1-p)^{n-k} p^k \\ +(1-p)^k p^{n-k}]$$

where $0 \leq m \leq 4$, and the p value is set to the same value used for figure 2. Note that p was not set to find the best model of rsequence (in terms of Root Mean Square, for example).

3.2. An alternative evaluation function

Starting from the considerations made above we wanted to formulate an evaluation function that awarded those predictors able to match the average accent pattern of human speakers and that was less sensible to n and m.

To satisfy these specifications we associated 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 emis-

	f1a	f1a-f2b-f3a-m1b-m2b-m3b	Δ(diff. between the first 2 colums)	ΔBaseline/ΔPredictor B
Baseline, OE(m=1)	46.88	73.17	26.29	-
Baseline, EE	48.88	69.54	20.66	-
Predictor B, OE(m=1)	75.34	97.56	22.22	1.18
Predictor B, EE	75.34	95.00	19.66	1.05

Table 2: *Comparison between OE and EE on predictor B and A (baseline).*

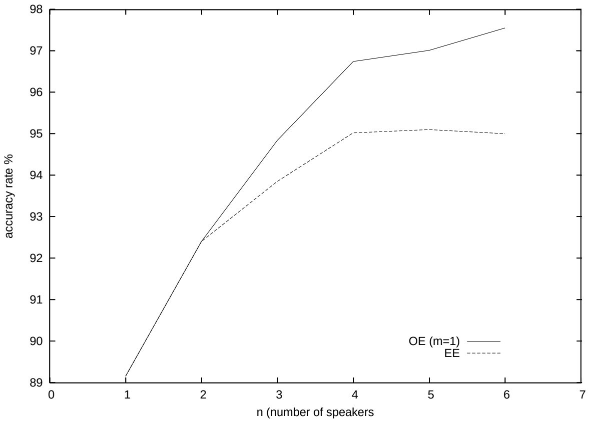

Figure 5: *Accuracy rates of Predictor B using OE and EE.*

sion is independent form the others.
From Information Theory ([7]) we know that the entropy of such a source is:

$$H = -\log(P(A))P(A) - \log(P(N))P(N) \quad (2)$$

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. If we associate the optionality of a word token with its entropy, and search for an evaluation function that is dependent on optionality, we can write the evaluation function for a single token as follows:

$$EE(w_i) = 1 - [(1 - P_t(pe_i))(1 - H_t(w_i))] \quad (3)$$

where $P_t(pe_i)$ is the probability that the predicted event pe_i is emitted by the test source and $H_t(w_i)$ is the entropy of the source. The overall EE is the sum of each $EE(w_i)$ divided by the total number of words.
The main novelty of EE is that intra-speaker optionality is no more simply considered as a binary quantity but as a gradient one.
Concerning the dependency on n and m, one of the practical advantages of EE is that we do not have to decide which the most appropriate value of m is, while regarding n we can see how EE is more stable than OE to n increase, if we suppose of having an infinite number of speakers. In that case, it is acceptable to assume a non-zero probability for each token of being assigned both pitch events, especially if we think that an error can be made by the speakers themselves or by the prosodic annotators. Both an all-accented and an all-non-accented predictors would score $OE(w_i) = 1$ per each token though neither of them would match the speakers average pitch pattern. Using EE both predictors would never reach the maximum score. This is an interesting characteristic of EE since usually a predictor performance is evaluated relatively to an all-accented or an all-non-accented baseline.
In order to provide some empirical evidence of the higher stability of EE we compared the two functions using different predictors. In figure 5 a predictor is evaluated on different values of n: for $n > 3$ the EE values are more stable than the OE values which keep on rising. Figure 5 shows the result for only one predictor evaluated over one out of 720 possible sequences of speakers. We carried out the same type of comparison using different predictors and different speaker sequences finding always the same kind of result. Table 2 reports the results of another type of comparison between EE and OE (with OE having $m = 1$). For both functions we computed the difference between the value obtained with $n = 1$ (first column) and that one with $n = 6$ (second column).
The table shows (third column) that for both measures the difference (Δ) between $n = 1$ and $n = 6$ obtained on the all-accented predictor is higher than our predictor, that means that the baseline increases more quickly than our predictor. Nevertheless, using EE, the increase of the baseline with respect to our predictor is slightly smaller: the fourth column of table 2 shows that when using EE the ratio between the Δ's of baseline and predictor (fourth row) is lower than that obtained using OE (third row). The choice of the speaker when $n = 1$ is not determinant since when a predictor is compared to a single speaker EE and OE assign the same score.

4. Intra-Speaker and Inter-Speaker Optionality

Until now we have seen how intra-speaker optionality can be taken into account when evaluating a pitch accent predictor assuming that the optionality occurring among speakers is the same optionality occurring within a single speaker (and consequently within a good predictor).

In order to explore to which extent this assumption is true, we compared two different predictors: a predictor trained on single speaker data (henceforth SSP) and a predictor trained on multi-speaker data (henceforth MSP) . Both training data consists of 8954 words. SSP was trained using a subset of the f2b section of the BURN corpus, whereas the MSP training set was built by grouping all the six speakers data of section p, r and t of the multi-speaker data, so the text read by the speakers (1293 words) and the values of the training features are repeated six

	f1a	f2b	f3a	m1b	m2b	m3b	All
SSP	76.15	83.2	82.93	87.26	82.93	84.01	93.87
MSP	75.34	82.93	83.74	89.16	82.66	84.82	95.00

Table 3: *Comparison between a predictor trained on single speaker data (SSP) and one trained on multi-speaker data (MSP).*

times (one for each speakers); as a consequence only the pitch accent values vary. The section j (369 words) was held out for testing both predictors. Both predictors were trained using the Classification and Regression Tree (CART [8]) available in the Edinburgh Speech Tools Library (Wagon CART [9]). We used training features that have been proven strictly correlated to prosodic prominence: part of speech (the MXPOST tagger [10] was used), logarithm of unigram and bigram of the word. Each example consisted of the feature values of a word and of the two words preceding and following it. Unigrams and bigrams were computed on a corpus of 17 million words (Herald news from 1998 to 2002) using the CMU toolkit for language modeling ([11]). Because of the smaller lexical variability of the multi-speaker data set we did not use lexical training features, like the accent ratio feature ([12]), that would have largely favored SSP. Both SSP and MSP were tested comparing their predictions with each one of the six speakers, and with all the six speakers at the same time using the EE evaluation function.
Looking at table 3, the most evident fact, when comparing the two predictors, is that their performances are very close. Surprisingly SSP performs slightly better than MSP when tested on three of the six speakers, whereas it is worse than MSP in the all-six-speaker evaluation. There results can be interpreted looking at a CART as a list of prediction rules: we can say, with a certain degree of approximation, that during the MSP training those rules that were sensitive to speakers, that is, appropriate for describing the pitch patterns of some speakers but not for those of the others speakers, were filtered out, so only the rules that assign the non-optional pitch events were successful. If the SSP performances are very close to the MSP ones we can conclude that, at least in our prediction model, the SSP has the same ability of the MSP to distinguish between intra-speaker optional and compulsory pitch events, but this is possible if the variability (with respect to training features strictly correlated to pitch accents) "seen" by the SSP during its training phase is very similar to the intra-speaker optionality seen by the MSP.
The Wagon CART provides, along with the predicted value, the probability of all the possible values (two, in our case) of the predicted variable. In the next section we compute the entropy of each prediction from the probabilities provided by Wagon and use this entropy as a training feature (henceforth called "uncertainty") to predict pitch accent optionality.

5. Predicting Intra-Speaker Optionality

Once we have formally defined intra-speaker optionality and shown the large overlap between intra and inter speaker optionality in our prediction model, we can try to predict optionality in order to improve the prosodic realization of unit selection TTS. In [13] it has been shown that including the pitch accent feature in the target cost function improves the quality of the unit selection speech synthesis. If we were able to associate to each predicted event its degree of optionality we would be able to tune the target cost associated with the pitch accent feature in accordance to the importance (optionality) of the pitch event. Informally, the less optional the pitch event is the more selective the unit selection module should be. This approach only considers the phonological aspect of a pitch event, that is its binary value accent/no-accent; optionality could be also correlated to the phonetic realization of pitch accents and this correlation could be used to improve prosodic modeling. However in this work we do not advance this possibility.
A predictor combining the prediction of the pitch event with the prediction of its correlated optionality could be evaluated using the following formula:

$$\begin{aligned} &EVA(w_i) = 1- \\ &\lambda[(1 - P_t(pe_i))(1 - H_t(w_i))(1 - H_p(w_i))] \\ &-(1 - \lambda)[H_t(w_i) - H_p(w_i)]^2 \end{aligned} \tag{4}$$

whith $0 \leq \lambda \leq 1$.
H_t and H_p are the actual and the predicted optionality respectively.
The first term of the sum in the squared parentheses evaluates the prediction of the pitch event taking into account how this event is considered optional by the predictor and how it actually is. The product of the predicted and the actual optionality guarantees a null error when at least one of the two optionalities is 1. The second term evaluates the optionality prediction. The two evaluation are weighted by the constant λ.
We tried to predict intra-speaker optionality training and testing the Wagon CART using again the multi-speaker section of the BURN corpus: 1293 words were used for training and 369 words hold out for testing.

	A	B	C	D
Otpionality	0	0.6500...	0.9182...	1

Table 4: *Entropy values given 6 speakers. Optionality values are associated to letters. A occurs when all the speakers agree, B when only one speaker disagrees, and so on.*

Unfortunately the data available were very small, so we have to consider the results we achieved still preliminary. The training features were the same used for training the pitch accent predictors (contextual features included) plus lexical form (only if the word occurred at least five times in the training set), distance (in number of words) from the closest punctuation mark form left and from right, and the "uncertainity" of the multi-speaker pitch-accent predictor. We thought that this last feature was not only an indicator of the approximation of the multi-speaker pitch accent predictor but also a quantity correlated to the intra-speaker optionality.

Given six speakers, there are only four possible values of optionality (table 4) for each word token. We found out that, in order to improve the learning phase, considering optionality as a categorical feature and associating to each optionality value a symbol, allowed us to achieve better results. The performances of our predictor were compared with an all-non-optional baseline, which assigns a zero-value to each token (this was also the most frequent optionality value). In table 5 we show the results

	ABCD	ABD	AD
Baseline	0.3066	0.3066	0.3066
Predictor	0.2718	0.2837	0.3066

Table 5: *Error rate in predicting optionality.*

when all the four optionality values were considered (ABCD) and when the number of values were reduced. For example, observing that the C and D values are very close we grouped them together (ABD). It is interesting to note that when we considered optionality as a binary feature by grouping all the non-zero values in a single symbol (D), we were not able to improve over the baseline.

In the training phase we used the Wagon "stepwise" option that only selects those training features that give a significant contribute in the learning phase. The "uncertainty" feature turned out to be the best one. Even using it as the only feature we achieved an improvement over the baseline. We also found out that if we substituted the uncertainty of the MSP with that of the SSP, the uncertainty feature was still the best one and we were still able to improve over the baseline.

6. Conclusion and Future Works

Our work has addressed some questions concerning intra-speaker disagreement and optionality in pitch accent placement: how "diffuse" is intra-speaker disagreement? How can we evaluate a pitch accent predictor on a multi-speaker testing data set? Is intra-speaker optionality predictable? Are intra-speaker and inter-speaker optionality the same thing with respect to our prediction model? How can we exploit optionality to improve unit selection text-to-speech synthesis?

We have shown the degree of intra-speaker optionality in read speech by analyzing six speakers and then we have proposed a new definition of intra-speaker optionality associating the concept of optionality to that of entropy. This mathematical definition allowed us to formulate a new evaluation function for evaluating pitch accent predictors which we proved to be more appropriate than the evaluation functions adopted in previous works. We then compared a predictor trained on a single speaker data with a predictor trained on multi-speaker data and from the high similarity of their predictions we inferred that a large overlap between inter and intra-speaker optionality exists. Supported by this result we suggested a simple strategy to improve the performances of a unit selection speech synthesis system that includes the pitch accent feature into its target cost features. Since this approach requires optionality be predictable, we tried to predict it and we achieved successful results. However we believe there is still room to improve our results and in the future we will try to improve them using larger data sets. Moreover in our experiment we only used training features that convey general properties of words. We believe that, since pitch accents have been proven to be prosodic correlates of the informativeness and significance of words (see [13], for example), the degree of optionality of a pitch accent is strongly correlated to the informative and significance status of the word the pitch accent is assigned to. Using POS, unigrams and bigrams we access only a part of that status, since we do not take into account the context in which words are and how their information status relates with it. In future work, we will consider linguistic features describing information structure (the contrast feature, for example) that have been proven being useful in detecting "meaningful" pitch accents [15] and evaluate our approach as part of a speech synthesis system.

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

[2] 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.

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

[4] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus". Technical Report ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, Boston, USA, 1995.

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

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

[7] C.E. Shannon, "A mathematical theory of communication". Bells System Technical Journal, 27:379-423 and 623-656, 1948.

[8] L. Breiman, J. Friedman, R. Ohlsen, and C.Stone, "Classification and regression trees", Wadsworth International Group, Belmont, USA , 1984.

[9] 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.

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

[11] P.R. Clarkson and R. Rosenfeld. "Statistical Language Modeling Using the CMU-Cambridge Toolkit", in Proc. ESCA Eurospeech 1997, Rhodes, Greece, 1997.

[12] A.Nenkova, J. Brenier, A. Kothari, S. Calhoun, L. Whitton, D. Beaver, D. Jurafsky, "To Memorize or to Predict: Prominence Labeling in Conversational Speech" in Proceedings of NAACL 2007, Rochester, USA, 2007.

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

[14] S. Pan, and K. McKeown, "Word informativeness and aoutomatic pitch accent modeling". Proc. of joint SIGDAT conference on empirical methods in natural language processing and very large corpora, 1999.

[15] S. Calhoun, "Information Structure and the Prosodic Structure of English: a Probabilistic Relationship", PhD thesis, University of Edinburgh, 2006.