OPTIONALITY IN EVALUATING PROSODY PREDICTION

Erwin Marsi

Tilburg University
ILK / Computational Linguistics and AI
Tilburg, The Netherlands
e.c.marsi@uvt.nl

ABSTRACT

This paper concerns the evaluation of prosody prediction at the symbolic level, in particular the locations of pitch accents and intonational boundaries. One evaluation method is to ask an expert to annotate text prosodically, and to compare the system's predictions with this reference. However, this ignores the issue of optionality: there is usually more than one acceptable way to place accents and boundaries. Therefore, predictions that do not match the reference are not necessarily wrong. We propose dealing with this issue by means of a 3-class annotation which includes a class for optional accents/boundaries. We show, in a prosody prediction experiment using a memory-based learner, that evaluating against a 3-class annotation derived from multiple independent 2-class annotations allows us to identify the real prediction errors and to better estimate the real performance. Next, it is shown that a 3-class annotation produced directly by a *single* annotator yields a reasonable approximation of the more expensive 3-class annotation derived from multiple annotations. Finally, the results of a larger scale experiment confirm our findings.

1. INTRODUCTION

Many speech synthesizers produce intonation in two steps. First, intonational events are predicted at the symbolic level. This involves placing pitch accent and phrase boundaries. Depending on the descriptive model, this may include specifying the particular type or strength of an accent or boundary. Second, this symbolic representation receives a phonetic realization in terms of F_0, segmental duration, and pause. No matter how good the phonetic realization, if the first step produced an ill-formed output, the spoken intonation will be judged negatively. The same goes for the reverse: even with perfect predictions at the symbolic level, the intonation produced can still be bad because of errors in the phonetic realization. Hence, it makes sense to evaluate both steps separately.

In this paper, we will be concerned with the evaluation of intonation at the symbolic level. There are basically two evaluation methods for this, roughly corresponding to speech perception and production. One is an experiment in which subjects are asked to judge the quality of the intonation for a representative set of examples with controlled prosody. In the case of lay subjects, who are not familiar with judging abstract prosodic categories, the stimuli must be synthetic or otherwise manipulated speech. Here, we run into the risk that errors in the phonetic realization distract subjects from judging the intonation proper. Alternatively, we can ask phonetically-trained experts to judge the symbolic representation directly [1]. Either way, running evaluation experiments is rather time consuming, in particular if we need quick diagnostic testing or tuning of data-driven approaches to prosody prediction. Moreover, there is no way of reusing the subject's judgments.

The second evaluation method is to measure how well the system's predictions match the actual behavior of human speakers. For this, we need to transcribe speech at the intonational level. Again, this is a time consuming activity, especially if we want to use several transcribers in order to increase the reliability of the transcription. Alternatively, we can ask subjects to annotate text prosodically . We have found that with an annotation tool providing feedback by means of speech synthesis and some amount of training, subjects can learn to indicate acceptable locations of pitch accents and major prosodic boundaries. It has been shown that a subject's prosodic annotation matches his/her actual prosodic realization when speaking reasonably well [2]. An advantage of this evaluation approach is that once obtained, either by transcription or annotation, the human reference can be reused for purpose of evaluation as often as we want to.

Ideal as this may seem, there is one important drawback: there is usually no unique solution. That is, for any utterance of reasonable length and complexity, there is more than one acceptable way to place accents and boundaries. This variability may stem from differences in interpretation, speech rate, personal preference, etc. Hence, a comparison of our system's output against a human reference is inherently asymmetrical. If the prediction matches the reference, it is correct, but if it does not match, this does not necessarily mean that it is incorrect. It may well be a perfectly acceptable alternative.

We can try to deal with this by transcribing many spoken versions of the same text or have several annotators annotate the same text independently, in the hope that we capture the full scope of variation. However, this would bring us back in the realm of expensive approaches to evaluation. In this paper, we propose another way to deal with the variability inherent in intonation. It stems from the intuition that some intonational events are optional whereas others are obligatory in nature, and that a human annotator is actually capable of considering multiple alternatives. Instead of a binary classification, we use an annotation in terms of optional and obligatory intonational events. Our aim is to show that this 3-class annotation leads to a more realistic estimation of the performance of a prosody prediction system, and that it offers a practical approximation of the more expensive method of using multiple annotations.

In the next Section, we describe an experiment in predicting the locations of accents and intonational boundaries, evaluating the output against multiple 2-class annotations and a single derived 3-class annotation. We argue in favor of the latter method, because it allows for detecting the real errors and gives better performance estimates. Section 3 compares a 3-class annotation derived from multiple annotation with a 3-class annotation by a single annotator. It is shown that a direct 3-class annotation forms a reasonable approximation of the more expensive derived 3-class annotation. Section 4 reports on a larger scale comparison of a 2-class versus a 3-class annotation, both produced by a single annotator. The last Section provides a summary and discussion of the results.

2. EXPERIMENTAL EVALUATION USING 2-CLASS AND 3-CLASS ANNOTATIONS

This experiment concerns predicting the locations of accents and intonational boundaries by means of memory-based learning. The experimental setting is described in more detail in [3]. The new issue here is that we compare an evaluation using a 2-class annotation with an evaluation using a 3-class annotation.

2.1. Training data

Our training data (the Prosit corpus) consists of 191 Dutch newspaper texts (55,192 words, excluding punctuation). All material was prosodically annotated (without overlap) by four different annotators, and corrected in a second stage (again without overlap) by two corrector-annotators. Annotators indicated the locations of accents and/or boundaries that they preferred. They used a custom annotation tool which allowed for feedback in the form of synthesized speech. In total, they assigned 22,227 accents (40% of all words) and 8627 breaks (16% of all word junctures). About half of the breaks (4601) were sentence-internal rather than sentence-final breaks.

In addition, all tokens were automatically annotated with shallow features of different categories (see [3] for details):

- orthographic: preceding and following punctuation symbols, presence of a diacritical accent
- syntactic: Part-of-Speech tag, Noun Phrase and Verb Phrase chunks
- informational: Information Content (IC), IC of bigrams, TF*IDF (a measure for how salient a word is within a document), and phrasometer (the summed log-likelihood of all n-grams the word form occurs in)
- positional: distance to previous occurrence of the same token, and normalized distance to start and end of the sentence

2.2. Test data

The test data is based on an independent corpus which was collected by Herwijnen & Terken to evaluate Dutch TTS systems [4]. It contains 2 newspaper texts (786 words), and 15 email messages (1133 words). All text material was independently annotated by 10 phonetically-trained experts for preferred accent and boundary locations (without any feedback by speech synthesis). Originally, four levels of boundary strength were distinguished, from ’0’ for ’no boundary’ up to ’3’ for ’strong boundary’. For our purposes, levels 2 and 3 were reduced to the class ’boundary’ and levels 0 and 1 to ’no boundary’. The testing material was also automatically annotated with the same features as for the training material.

From these 10 expert annotations, we then derived a 3-class annotation in the following way:

- if all 10 experts assigned an accent/boundary, then it is obligatory;
- if none of the experts assigned an accent/boundary, then it is impossible;
- in all other cases (that is, if the experts disagree), it is optional.

For the of purpose of evaluation, we interpret *optional* here as meaning: whether an accent or no accent is predicted, either way is correct. Therefore, in calculating the difference between a 3-class annotation and a 2-class prediction, we will simply *ignore* all optional cases. As a consequence of this approach and the way that the 3-class annotation was derived, none of the experts differs from the 3-class annotation and thus all score 100% on whatever evaluation metric is used (accuracy, F-score, kappa).

	Accents :		*Boundaries* :	
Expert :	*News* :	*Email* :	*News* :	*Email* :
1	62.0	56.1	91.7	76.0
2	82.0	79.2	91.1	86.8
3	84.7	83.5	91.0	85.7
4	82.3	54.4	84.5	82.4
5	76.3	81.2	67.6	77.3
6	77.6	80.1	86.5	85.0
7	74.1	78.2	84.6	84.2
8	85.7	78.4	90.1	78.1
9	73.2	83.6	89.5	80.7
10	81.4	82.1	87.3	83.7
av:	77.9	75.7	86.4	82.0
sd:	11.1	11.0	7.13	3.77

Table 1. F-scores for obligatory accent and boundary predictions by TiMBL on news and email texts calculated against the 10 expert annotations

2.3. Learning

Learning was performed with TiMBL [5], our enhanced implementation of the k-NN classification algorithm. In previous work [3], we showed that TiMBL performs slightly, but significantly, better than CART, the often used decision tree approach to prosodic classification [6]. In order to model a token's context, instances were created by applying a window. For accent, we used a 3-1-3 window, which means that for each feature we also included its values in the preceding three and following three instances. Likewise, a 1-1-1 window was used during boundary prediction. The positional features were excluded from windowing. Furthermore, TiMBL's parameters were optimized for the task on the training material as described in [3].

2.4. Results

Table 1 presents the F-scores of accent and boundary predictions on news and email texts calculated against each of the 10 expert annotations. The standard deviations are rather high, indicating a substantial amount of variation among the ten expert annotations. As a matter of fact, calculating the F-score of each expert against the other nine, and taking the average of these scores, gives 80.4 for news accents, 74.2 for mail accents, 86.0 for news boundaries, and 81.9 for email boundaries.

Evaluating the predictions against the derived 3-class annotation yields an F-score of 92.2 for the news accents, 90.4 for the email accents, 94.4 for the news boundaries, and 93.9 for the email boundaries. Clearly, these are huge improvements (of 12 up to 16 points) over the average F-scores calculated against the 10 experts annotations.

2.5. Discussion

Note that this is *not* a trivial result: decreasing the test set by *randomly* ignoring a number of instances will normally not improve the F-score. Of course, our sample is not random, but corresponds to the optional cases, which we defined as the ones the experts disagreed upon. The results therefore reveal that many of the seemingly incorrect predictions of our classifier are in fact in agreement with at least one of the experts. In other words, they involve optional rather than obligatory cases. Moreover, the classifiers' crucial predictions (i.e. where an accent/boundary is either obligatory or impossible) are actually better than suggested by the average F-score from the first evaluation method. Our conclusion so far is that a 3-class annotation derived from multiple independent annotations allows us to detect the real prediction errors and leads to more representative performance measures.

3. COMPARING DERIVED AND DIRECT 3-CLASS ANNOTATIONS

Despite the advantage of evaluating against a 3-class annotation, deriving it from 2-class annotations produced by ten human experts is rather expensive, and seems infeasible for substantial amounts of text. Alternatively, a single human expert may attempt to produce a 3-class annotation directly. In this section, we address the question to what extent such a direct 3-class annotation is equivalent to a derived 3-class annotation.

The news and email texts were annotated by a single annotator (not one of the ten experts discussed earlier) for obligatory and optional accents/boundaries. The remaining unannotated tokens were assumed to belong to the third class (i.e. impossible). A custom annotation tool was used which allowed for feedback by synthesized speech from the Nextens speech synthesizer.[1] Utterances could be synthesized with only the obligatory accents/boundaries, or with both the obligatory and optional ones. In the latter case, the speech rate was slightly decreased. The annotator was aware that the feedback is helpful, but not to be blindly trusted (for instance, it did not produce a proper intonation contour for questions). Her annotation was reviewed by the author and some minor revisions were made.

Table 2 shows the correspondence between the direct 3-class annotation and the number of times that n experts assigned an accent/boundary, for both the news and email texts. It is clear that overall obligatory accents/boundaries tend to correspond to most or all of the experts assigning an accent/boundary, whereas impossible accents/boundaries correspond to few or none of the experts assigning an ac-

[1] The homepage of the Nextens project is `http://nextens.uvt.nl`

		0:	1:	2:	3:	4:	5:	6:	7:	8:	9:	10:
$News$	$*$:	1	9	6	7	9	9	19	19	35	54	115
acc :	$+$:	6	10	8	9	10	11	5	8	4	2	1
	$-$:	383	23	8	8	1	3	1	2	0	0	0
	tot :	390	42	22	24	20	23	25	29	39	56	116
$Email$	$*$:	1	3	4	7	5	13	31	48	86	106	81
acc :	$+$:	5	12	7	8	13	18	13	14	10	7	2
	$-$:	523	62	22	19	6	6	1	0	1	0	0
	tot :	529	77	33	34	24	37	45	62	97	113	83
$News$	$\#$:	0	1	0	1	1	2	3	2	8	29	43
bnd :	$\|$:	4	2	0	0	0	0	0	0	0	0	0
	$-$:	658	19	7	2	2	1	0	1	0	0	0
	tot :	662	22	7	3	3	3	3	3	8	29	43
$Email$	$\#$:	0	3	2	3	0	5	7	12	6	27	69
bnd :	$\|$:	1	0	2	0	1	0	1	0	0	0	0
	$-$:	924	36	10	12	11	2	0	0	0	0	0
	tot :	925	39	14	15	12	7	8	12	6	27	69

Table 2. Correspondence between the direct 3-class annotation by a single annotator and the number of times that n experts assigned an accent/boundary, where $*$ = obligatory accent , $+$ = optional accent, $\#$ = obligatory boundary, $\|$ = optional boundary, $-$ = impossible accent/boundary, and tot = total no of times that n experts agreed

cent/boundary. Optional accents show no such tendencies. The pattern is less clear for optional boundaries, if only because there a so few of them.

In addition to looking at the raw data, we can measure the fit between both of the 3-class annotations in terms of F-scores. It is possible to calculate F-scores on the three classes, assuming that the derived annotation gives the real classes and the direct annotation gives the predicted classes. However, since we cannot compare these with any of our previous figures, these are hard to interpret. Instead, Table 3 presents the scores for obligatory accents and boundaries in the ten expert annotations evaluated against the direct 3-class annotation. This is the type of evaluation in which the optional real classes are ignored. The reasoning behind this is the following: given that each of the expert annotations scores 100 evaluated against the *derived* annotation, if they score comparably high evaluated against the *direct* annotation, then the direct annotation must be very similar to the derived annotation. Unfortunately, the scores are not nearly close to 100, so we have to conclude that the direct annotation is similar though not fully equivalent to the derived one.

Finally, the predictions from the experiment described in the previous Section were evaluated against the direct 3-class annotation. The F-scores, together with those obtained earlier for evaluating against the 10 experts separately and against the derived 3-class annotation, are shown in Table 4.

	$Accents$:		$Boundaries$:	
$Expert$:	$News$:	$Email$:	$News$:	$Email$:
1	72.5	61.5	95.5	81.2
2	89.1	88.2	93.8	89.3
3	94.5	93.4	93.6	93.0
4	91.4	65.3	89.7	89.1
5	85.3	87.7	71.7	85.4
6	89.4	91.9	90.1	84.3
7	86.0	90.7	90.4	90.6
8	91.4	90.6	95.0	89.6
9	82.8	89.1	93.5	80.0
10	90.4	93.4	95.7	88.4
av:	87.3	85.2	90.9	87.1
sd:	6.2	11.7	7.1	4.2

Table 3. F-scores on obligatory accent and boundary assigned by the 10 experts evaluated against the direct 3-class annotation

	$Accents$:		$Boundaries$:	
$Annotation$:	$News$:	$Email$:	$News$:	$Email$:
10 expert:	77.9	75.7	86.4	82.0
derived 3-class:	92.2	90.4	94.4	93.9
direct 3-class:	90.8	89.0	89.6	86.8

Table 4. F-scores of accent and boundary predictions by TiMBL on news and email texts calculated (1) against the 10 expert annotations, (2) against the derived 3-class annotation, and (3) against the direct 3-class annotation

Consistent with the previous findings, the results suggest that evaluating against a direct annotation produces higher estimates of F-scores, in particular for accents, though not as much as evaluating against the derived annotation.

To sum up, we have shown that for practical purposes a direct 3-class annotation gives a reasonable approximation of the more expensive derived 3-class annotation.

4. SCALING UP

In this Section we report on a larger scale experiment, using about three times as many tokens, directly annotated in three classes.

The Prosit corpus originally used for training (cf. Section 2) was divided in 90% training material (49.663 tokens) and 10% test material (5529 tokens). In addition, the test material was directly annotated in three classes using the same procedure as described in the previous Section and by the same annotator. This 3-class annotation was produced from scratch, without looking at the 2-class annotation. Learning was again carried out with TiMBL, using the same feature windowing and the same optimized parameter settings.

To ensure that our results are more than trivial, we constructed two baselines. The baseline for accent placement is based on the content versus function word distinction, commonly employed in TTS systems [7]. It is constructed by accenting only content words, leaving all function words (determiners, prepositions, conjunctions/complementisers and auxiliaries) unaccented. The required word class information is obtained from the POS tags. The baseline for breaks relies solely on punctuation. A break is inserted after any sequence of punctuation symbols containing one or more characters from the set {,!?:;()}. It should be noted that both baselines are simple rule-based algorithms that have been manually optimized for the current material. They perform well above chance level, and pose a serious challenge to any ML approach.

Table 5 summarizes the F-scores obtained for evaluating against the 2-class annotation and against the 3-class annotation. As before, instances for which the real class is optional are ignored. As already reported in [3], the boundary predictions are only marginally better than the baseline, primarily because the precision of the baseline is nearly perfect. Apart from this, the results are consistent with those reported above. Our interpretation is that again quite a lot of the seemingly incorrect predictions in fact concern optional rather than obligatory cases. At places where the decision really matters (i.e. where an accent/boundary is either obligatory or impossible), the predictions are actually better than suggested by evaluating against the 2-class annotation.

The observation that the gain is rather large for boundaries may be attributed to the effect of sentence-final boundaries. As is common in reporting scores on phrase break prediction (e.g. [7]), breaks at the end of a sentence are included in the count. However, if end-of-sentence detection errors are manually corrected – as in most work, including ours – then predicting sentence-final breaks becomes a trivial task, and just boosts the score. Now moving to a 3-class annotation means that some of the boundaries that were originally obligatory will become optional, so in effect the number of real, sentence-internal boundaries decreases. Therefore, the contribution of the sentence-final boundaries to the score becomes bigger, and since these are always correct, the score becomes better. Looking at our data, the 2-class annotation has 860 boundaries, 450 of which are sentence-internal and 410 sentence-final. The 3-class annotation has only 746 obligatory boundaries, consisting of 336 sentence-internal and 410 sentence-final boundaries. As expected, the ratio of sentence-internal versus sentence-final boundaries has become smaller, so one may think that explains most of the gain.

Despite this seemingly convincing argument, this turns out to *not* be the case. Ignoring the sentence-final breaks, our score on the sentence-internal breaks in the 2-class annotation is 72.9, whereas that on the 3-class annotations is 83.3. This suggests that the gain should really be attributed to the addition of the optional class.

$Predictor:$	$Annotation:$	$Accents:$	$Boundaries:$
$Baseline:$	2-class	80.3	86.1
	3-class	81.4	92.4
$TiMBL:$	2-class	85.5	87.1
	3-class	88.7	92.8

Table 5. F-scores for obligatory accent and boundary predictions by the baseline and by TiMBL, evaluated against the direct 2-class and 3-class annotations

5. SUMMARY AND DISCUSSION

To sum up, we have proposed dealing with optionality in evaluating prosody prediction by means of a 3-class annotation which includes a class for optional accents/boundaries. We have demonstrated that evaluating against a 3-class annotation derived from multiple independent 2-class annotations allows us to identify the real prediction errors and to better estimate the real performance. It was shown that a 3-class annotation produced directly by a *single* annotator yields a reasonable approximation of the more expensive 3-class annotation derived from multiple annotations. Finally, the results of a larger scale experiment were consistent with our findings.

Using prosodic *annotation* of text rather than a prosodic *transcription* of speech is not a new idea. For instance, [8] reports on an evaluation of a rule-based TTS system for Dutch by comparing the predicted accents with those prescribed by a number of experts. [2] show that this is an acceptable strategy, because speakers can reliably predict what prosodic structure they will realize when reading text aloud.

The fact that there is usually more than one acceptable way to place accents and boundaries is well-established in the linguistic literature, and has been acknowledged by many researchers working on automatic prediction, including those using transcribed prosody [6, 9]. [10] show experimentally that many of the predicted phrase boundaries that do not match the reference are in fact judged as fully acceptable by human listeners. [1] describe an evaluation in which experts judge the symbolic output of a prosody prediction system. Although this solves the issue of optionality, in the same way as judging the appropriateness of synthesized speech does, it prevents the reuse of subjects' judgments. This makes it unsuitable for the testing and tuning typically required in data-driven approaches to prosody prediction [11, 12, 13, 14]. Attempts to address this issue by introducing a third optional class can be found in [15, 11, 13]. Interestingly, [11] also show that this significantly improves the performance on phrase break prediction using decision trees. However, to the best of our knowledge, it has not been shown before that a single direct annotation including an optional class is highly similar to an annotation derived from multiple annotations without an optional class. Also, it has not been reported before that this improves performance on accent placement in a machine learning context.

A possible drawback of a 3-class as opposed to a 2-class annotation is that it takes more time, because the annotator has to consider more classes and alternatives. However, it is probably still faster than transcription (for instance, annotating the 1119 words of news and email text took 14 hours). Moreover, the annotation speed will increase once we are capable of predicting the three classes automatically with a reasonable accuracy, and the annotator only has to make corrections. Apart from time considerations, it is also worth mentioning that the 3-class annotation is more satisfactory to the annotator, because it reduces the feeling that one is forced to make arbitrary choices, as is often the case with the 2-class annotation.

Another potential problem is that choices may be context-dependent [2]. For example, sometimes a two-word phrase must have either its first word accented or its second word, but not both at the same time. Such cases cannot be expressed in our 3-class annotation scheme. However, they are rare, at least in our experience with annotation. An interesting solution is to produce graphs representing all possible annotation paths [16].

Our plans for future work include *training* with a 3-class annotation. [11] report that removing the optional instances from the training material improves performance on phrase break prediction, because it forces the classifier to make clear decisions. Besides reproducing similar results, we hope to prove that this is also the case for accent prediction.

Acknowledgements The Prosit corpus was built together with my colleagues Antal van den Bosch, Walter Daelemans, Véronique Hoste, and Martin Reynaert as part of the Prosit project (`http://ilk.uvt.nl/prosit`). We are grateful to Olga van Herwijnen and Jacques Terken for permitting the use of their TTS evaluation corpus, and to Corina Koolen for the 3-class annotation. Many thanks to all of them for useful comments on the paper.

6. REFERENCES

[1] A. Monaghan and D.R. Ladd, “Symbolic output as the basis for evaluating intonation in text-to-speech systems,” *Speech Communication*, vol. 9, no. 4, pp. 305–314, 1990.

[2] O.M. van Herwijnen and J.M.B. Terken, “Do speakers realize the prosodic structure they say they do?,” in *Proceedings Eurospeech 2001 Scandinavia*, 2001, vol. 2, pp. 959–962.

[3] E. Marsi, G.J. Busser, W. Daelemans, V. Hoste, M. Reynaert, and A. van den Bosch, “Learning to predict pitch accents and prosodic boundaries in Dutch,” in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 2003, pp. 489–496.

[4] O. van Herwijnen and J.M.B. Terken, “Evaluation of PROS-3 for the assignment of prosodic structure, compared to assignment by human experts,” in *Proceedings Eurospeech 2001 Scandinavia*, 2001, vol. 1, pp. 529–532.

[5] W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch, “TiMBL: Tilburg Memory Based Learner, version 5.0, reference manual,” Tech. Rep. ILK 03-10, Induction of Linguistic Knowledge, Tilburg University, 2001.

[6] J. Hirschberg, “Pitch accent in context: Predicting intonational prominence from text,” *Artificial Intelligence*, vol. 63, pp. 305–340, 1993.

[7] P. Taylor and A. Black, “Assigning phrase breaks from part-of-speech sequences,” *Computer Speech and Language*, vol. 12, pp. 99–117, 1998.

[8] R. van Bezooijen and L.C.W. Pols, “Evaluation of a sentence accentuation algorithm for Dutch text-to-speech system,” in *European Conference on Speech Communication and Technology*, Edinburgh, 1989, vol. 1, pp. 218–212.

[9] A.W. Black, “Comparison of algorithms for predicting pitch accent placement in English speech synthesis,” in *Proceedings of the Spring Meeting of the Acoustical Society of Japan*, 1995.

[10] M.C. Viana, L.C. Oliveira, and A.I. Mata, “Prosodic phrasing: Machine and human evaluation,” in *Proceedings 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire, Scotland, 2001.

[11] P. Koehn, S. Abney, J. Hirschberg, and M. Collins, “Improving intonational phrasing with syntactic information,” in *ICASSP*, 2000, pp. 1289–1290.

[12] G.J. Busser, W. Daelemans, and A van den Bosch, “Predicting phrase breaks with memory-based learning,” in *Proceedings 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire Scotland, 2001.

[13] E. Marsi, *Intonation in Spoken Language Generation*, LOT, Utrecht, 2001.

[14] E. Marsi, G.J. Busser, W. Daelemans, V. Hoste, M. Reynaert, and A. van den Bosch, “Combining information sources for memory-based pitch accent placement,” in *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, 2002, pp. 1273–1276.

[15] E. Marsi, “Automatic evaluation of intonational phrasing algorithms for Dutch,” in *Nordic Prosody, Proceedings of the VIIth Conference*, S. Werner, Ed., pp. 171–194. Peter Lang, Berlin, 1998.

[16] I. Bulyko and M. Ostendorf, “A bootstrapping approach to automating prosodic annotation for constrained domain synthesis,” in *Proceedings of the IEEE Workshop on Speech Synthesis*, 2002.