Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Integrating sequence information in the audio-visual detection of word prominence in a human-machine interaction scenario

Andrea Schnall, Martin Ernst Heckmann

2014

Preprint:

This is an accepted article published in INTERSPEECH. The final authenticated version is available online at: https://doi.org/[DOI not available]

Integrating Sequence Information in the Audio-visual Detection of Word Prominence in a Human-machine Interaction Scenario

Andrea Schnall[1,2]*, Martin Heckmann*[2]

[1] TU Darmstadt - Control Methods and Robotics
Holzhofallee 38, 64295 Darmstadt, Germany
[2] Honda Research Institute Europe GmbH
Carl-Legien-Str. 30, 63073 Offenbach/Main, Germany
aschnall@rmr.tu-darmstadt.de, martin.heckmann@honda-ri.de

Abstract

Modifying the articulatory parameters to raise the prominence of a segment of an utterance (hyperarticulating) is usually accompanied by a reduction of these parameters (hypoarticulation) for the neighboring segments. In this paper we investigate different approaches for the automatic labeling of the prominence of words. In particular, we investigate how the information in the sequence can be used. During the recording of the underlying audio-visual database, the subjects were asked to make corrections for a misunderstanding of a single word of the system by using prosodic cues only. We extracted an extensive range of features from the audio and visual channel. For the classification of word prominence we compare two algorithms. On the one hand SVM, a local classifier, on the other hand a classifier based on a sequential model, linear chain Conditional Random Fields (CRF). Both were trained on different context regions. For the CRF the whole sentence is used as a word sequence for training and testing. Overall we show that sequence models such as CRF, which performs best in our experiment, are suited for prominence detection and, furthermore, that the neighboring words contain information which further improves the detection.

Index Terms: prosody, prominence, audio-visual, sequence information

1. Introduction

The determination of the prominence of a word plays an important role in human communication. By strongly increasing the prominence of a word, for example, a correction can be indicated [1, 2]. Using this information in human-machine interaction systems is nevertheless a challenging task. Consequently, so far it has been rarely used in spoken dialog systems [3, 4]. The reason for this might be that prosodic cues are difficult to extract due to their variation depending on the speaker.

In human-machine communication misunderstandings are still quite frequent. By correcting these misunderstandings the error rate even increases [5, 2, 6]. In a lot of systems it is not possible or useful to use specific cue phrases, because of the lexical and syntactic constraints on human-computer dialog; therefore, the utterance is often repeated by emphasizing the correcting word [7]. By detecting prominent words, both presence and location of a correction can be determined [7].

In our previous work, we investigated the audio-visual discrimination between prominent and non-prominent words [8]. In this paper we compare two different machine learning approaches, namely Support Vector Machines (SVM) and Conditional Random Fields (CRF). Prior work (SVM:[9], CRF:[10, 11]) used these algorithms for pitch accent detection. Hence we expect them to be well suited for the detection of prominent words. For the pitch accent classification the authors used non acoustic features [10], i.e. syntactic, phonological, or acoustic features in combination with text-based features [11]. Our case of prominence detection differs from these approaches. As the subjects in our database were ask to correct misunderstandings by using prosodic cues, there is no correlation between prominence and text-based features. Therefore we use only acoustic features. Despite their widespread use in recent years for speech recognition, there is only one comparison of CRF with a local classifier that uses text-based features [11].

Additionally, we analyze the performance for features based on single words compared with the performance of features based on the knowledge of the neighboring words. This is based on the hypothesis that emphasizing a word by hyperarticulating leads to a hypoarticulation of the neighboring words [12, 13]. Also, other authors showed the value of incorporating context for pitch accent detection [14, 15, 9]. We therefore investigated - for the task of classifying a word - how including also the features of the previous and following words changed performance. We compared the classification of individual words, SVM, with a method for sequence labeling, CRF.

In the second section we will present the database used, followed by a short description of the features. In section 4 the different methods we applied are described, followed by results in section 5 and a discussion in section 6. In the final chapter we give a conclusion.

2. Dataset

The audio-visual data was recorded as a Wizard-of-Oz experiment in a small cartoon game [16]. The subjects had to give spoken advises, in a simple grammar, to a computer. Some of the words were misunderstood by the machine, which was visually shown. Then the subjects had to correct the sentence by repeating them and using prosodic cues to emphasize the misunderstood word. Correction cue phrases like 'no I said' were not allowed. This procedure should lead to a rather natural use of prosody with the corrected words marked as highly prominent.

For the experiments a subset of 8 subjects, male and female, speaking either British or American English as native language or being either bilingual British English/German or American English/German were recorded. Over 2500 utterances with 4

to 5 words each where used for evaluation. Their speech was recorded at 48 kHz and later downsampled to 16 kHz. For the video images a CCD camera with a resolution of 1280x1024 pixel and a frame rate of 25Hz was used. A speech recognition system trained on the Grid Corpus [17] was used for forced alignment.

Three annotators labeled the prominence of each word of the database on a 4 level scale (0-3). Based on this labeling we built a binary gold standard annotation. A word was annotated as prominent if the mean annotation of all annotators was above 1.5. During the annotation we saw that on the one hand, not all speakers consistently used prominence to highlight the corrected word and, on the other hand, that the annotators were for some speakers not able to come to a consensus on the prominence of the words. Therefore, from the originally 16 speakers we retained 8 which showed an agreement measured with Fleiss Kappa of more than 0.55 between all annotators (0.41-0.6 is usually considered as moderate agreement).

3. Features

In this section, we will shortly describe the features used - respectively the feature-combinations - before we describe the employed classifier and discuss the results.
From the audio-channel we extracted: the energy (en) of a word relative to the mean of the utterance, the duration of a word (D_W) as well as the gap length (D_G) before and after the word as determined from the forced alignment, the spectral emphasis (SE), i.e. the difference between the overall intensity and the intensity in a dynamically low-pass-filtered signal with a cut-off frequency of $1.5 f_0$ [18] and RASTA-PLP features (R) [19]. Furthermore, we calculated the mean fundamental frequency relative to the mean of the utterance $\left(\bar{f}_0\right)$, the maximal fundamental frequency $\left(f_0^M\right)$ and the maximal curvature of the fundamental curve $\left(f_0^C\right)$ [8].

From the visual channel, the region around the mouth is detected with the use of a nose tracking. Then a two-dimensional Discrete Cosine Transform (DCT) is calculated. Of these coefficients the 50 with the highest energy are extracted [8]. From all features the mean value for each word was calculated (except for the duration). The beginning and the end of the word is obtained from the forced alignment.
We combined this features in three different variations:

(a) prosodic features: $en - D_W - D_G - Se - \bar{f}_0 - f_0^M - f_0^C$

(b) prosodic features + RASTA-PLP:
$en - D_W - D_G - SE - \bar{f}_0 - f_0^M - f_0^C - R$

(c) prosodic features + RASTA-PLP + visual features:
$en - D_W - D_G - SE - \bar{f}_0 - f_0^M - f_0^C - R - DCT$

4. Methods

For classification a lot of different algorithms are available. A very popular method, the SVM, has already been used for this task [16]. For related problems in speech processing, graphical methods, such as the CRF have often proven to be a good choice [20], especially for sequence labeling problems, such as part-of-speech tagging. In this paper, we investigate their suitability to the task of word prominence detection.

As described before, modifying the articulatory parameters to raise the prominence of a segment of an utterance, is usually accompanied by a modification, e.g. a reduction of these parameters, for the neighboring segments [12]. This is why the prosodic features of the neighboring words should also change and be useful for classification. Therefore, in addition to using sequence labeling techniques, we also investigate if the word prominence detection can be improved by using features spanning over several words.

4.1. Support Vector Machines

For classification with support vector machines (SVM) we used the libsvm [21] with a radial basis function kernel. For the classification, the data was normalized to the range [-1,1]. For the parameter C, the penalty parameter of the error term, and γ, the variance scaling factor of the basis function, a grid search on the whole dataset was performed for every feature combination. Then we divided our data in a training set, containing 75% of the data, and a test set, containing the remaining 25% of the data, using 30 fold cross validation. The training and testing was done for every speaker independently. As features we used the three different combinations described in section 3.

We then extended the SVM features with the information of the neighboring segments. As mentioned before, the surrounding words are potentially modified, so they are different from the other non-prominent words as well as from the prominent words. Therefore, a combination of the features of the actual word and those of the neighbors should contain more information on the prominence of the central word than the single features of the actual word. The SVM were again trained as described above with the new context-features. Additionally, we did an experiment where we used only the mean/max fundamental frequency and the energy of the neighboring words.

4.2. Conditional Random Fields

In order to use the sequential information in the sentences, we applied conditional random fields (CRF) [22]. CRF is a discriminative, undirected probabilistic graphical model. We used the special case of a linear chain for modeling our utterances. A conditional model specifies the probabilities of possible label sequences given an observation sequence. They have proven to be a useful tool for speech recognition, especially for part of speech tagging [20, 11]. CRF is a model that allows for combination of local information to predict a global probabilistic model over sequences. In our case the whole utterance is used as a word sequence, where the features of every word correspond to the observation sequence and the labels of the words (prominent or not prominent) correspond to the label sequence, for training and testing.

For implementation we used the Graphical Models in Mallet [23], extended for continuous features. Again, as with SVM the training was done for every speaker independently and tested with varying context of neighboring words.

5. Results

Compared to the detection of pitch accents, prominent words occur less often. While e.g. in [10] the classes are nearly balanced (approx. 60% unaccented), in our database, there are approx. 8 times more non-prominent words than prominent words. Because of this unbalanced database, a comparison of the accuracy is not suitable; a preference of non-prominent words would always lead to very good accuracy but would not provide an objective evaluation. Therefore, we use the F1-score for evaluat-

Table 1: *F1-score and UA for the different context regions and three different feature combinations for classification with SVM, one time with context of all feature types and one time with context of only the energy and the fundamental frequency, CRF and CRF trained on 3 label classed (but evaluated for prominent/not-prominent). Bold values indicate the best result per method.*

region	1	2	3	4	5	6	7
SVM, context-regions only for the energy and fundamental frequency [F1 (UA)]							
a	57.33 (77.56)	64.96 (81.63)	64.65 (80.17)	66.65 (80.20)	66.05 (81.95)	66.66 (81.58)	68.26 (83.40)
b	70.08 (81.28)	71.55 (82.62)	72.51 (83.59)	72.51 (83.59)	71.15 (82.26)	70.08 (82.84)	**72.89 (83.99)**
c	69.13 (81.59)	69.90 (83.06)	69.88 (82.43)	71.08 (83.52)	70.66 (83.44)	70.44 (83.14)	71.65 (83.91)
SVM, [F1 (UA)]							
a	57.33 (77.56)	67.25 (82.32)	66.73 (82.29)	70.19 (83.08)	70.33 (82.31)	70.29 (82.35)	72.20 (84.39)
b	70.08 (81.28)	72.24 (84.22)	72.36 (84.68)	**74.46 (84.26)**	72.98 (83.56)	70.04 (82.04)	72.43 (83.35)
c	69.13 (81.59)	70.13 (82.01)	71.09 (83.01)	70.61 (82.41)	69.61 (81.72)	68.98 (81.39)	68.94 (80.38)
CRF, [F1 (UA)]							
a	55.98 (72.39)	68.72 (79.64)	75.07 (84.23)	76.65 (85.39)	72.09 (82.07)	77.10 (85.90)	**77.65 (86.27)**
b	65.86 (78.88)	66.36 (80.03)	74.95 (84.90)	74.11 (84.34)	66.91 (80.34)	73.91 (84.13)	73.03 (83.74)
c	64.97 (78.90)	64.70 (78.99)	73.30 (83.93)	72.42 (83.41)	65.25 (79.14)	73.25 (83.71)	71.90 (82.92)
CRF, trained on 3 label classes [F1 (UA)]							
a	58.98 (75.15)			77.60 (86.44)			**78.46 (86.90)**
b	65.07 (79.26)			73.53 (83.98)			74.61 (84.42)

ing the performance of the prominent words:

$$\text{precision} = \frac{tp}{tp + fp}, \quad (1)$$

$$\text{recall} = \frac{tp}{tp + fn}, \quad (2)$$

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \quad (3)$$

and the unweighted accuracy (UA):

$$\text{specificity} = \frac{tn}{tn + fp}, \quad (4)$$

$$\text{unweighted accuracy} = \frac{\text{recall} + \text{specificity}}{2}, \quad (5)$$

with tp = true positive, tn = true negative, fp = false positive and fn = false negative.

This section presents all results of the F1-score regarding the detection of prominent words and the unweighted accuracy of non-prominent and prominent words, using all words of the utterances (4 to 5 words). For the classification of every word, additionally to its own features, a different amount of context was used for the classification as well. The regions are:

- region 1: single features of the actual word
- region 2: features of the actual and the previous word
- region 3: features of the actual and the following word
- region 4: triple features (actual + previous + following word)
- region 5: features of the actual and two previous words
- region 6: features of the actual and two following words
- region 7: quintuple features (actual + 2 previous + 2 following words)

Table 1 shows the results for SVM and CRF for three different feature combinations (a), (b), (c) as described in section 3 and the different regions, averaged over the 8 speakers. For SVM, one time all feature types of the neighbor words were used, and one time only the energy and the mean/max fundamental frequency were used as in [15] for calculating the z-score of the neighborhood region. For both variations and all combinations, the incorporation of the context information improves the results. The highest increase of F1-score, from using the context of all feature types, can be seen for the feature combination (a) from 57.33% (UA: 77.56%) to 72.20% (UA: 84.39%) when using the quintuple features. For SVM, the results when using the previous or the following word were quite similar. But results indicate that using one previous and one following word leads to better results than either two previous or two following words. For feature combination (a) the quintuple feature worked best while for the feature combinations (b) and (c) the triple features worked best. Comparing this with the reduced context for SVM, the results show that these neighborhood features contain already a lot of information, but also that it is still useful to also consider the other features of the neighborhood for an even better result at least for feature combination (a), the prosodic features, and (b), the prosodic features and the RASTA-PLP, while for feature combination (c), the prosodic features and the RASTA-PLP and the visual features, the reduced features performed better.

The lower part of table 1 shows the results for CRF for the different regions, averaged over the 8 speakers. As for SVM, we can observe for CRF that the incorporation of the context information improves the results. The F1-score for the prosodic features (a) improves from 55.98% (UA: 72.39%) to 77.65% (86.72%) with quintuple features. While the first result (single features) is worse than the one we obtained for SVM, the results of CRF when using context is better than the results of SVM when using context. Using the CRF for classification with the region 2 and 3 (respectively 5 and 6) leads to very different results. The following word seems to contain much more information than the previous one. Still the combination with both leads to the best results. But again, for the combination (a), the quintuple features are the best, while for the feature combination (b), the triple features are the best.

Based on the hypothesis that emphasizing a word by hyperarticulating leads to a hypoarticulation of the neighboring words, we assign an extra class label for the previous and the

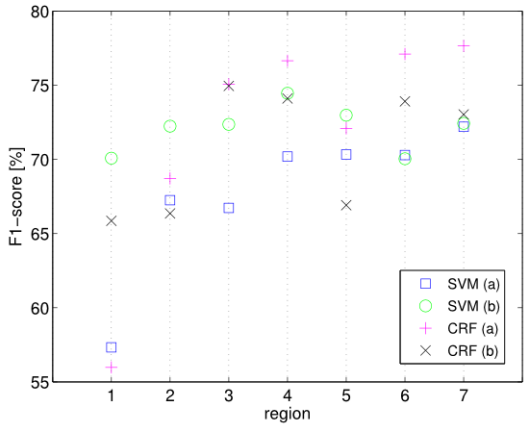


Figure 1: *Comparison of the F1-score of SVM and CRF, regarding the different context regions and feature combinations (a) and (b).*

following words of the prominent words. By treating the two not prominent classes as one class for the evaluation, the F1-score and the unweighted accuracy could be calculated as in the other experiments. The results for CRF using this new annotation for the single, triple, and quintuple regions are shown on the bottom of table 1. For both feature combination (a) and (b) the performance improved for this new annotation.

A direct comparison of SVM and CRF regarding feature combinations (a) and (b) can be seen in figure 1. The results for using single features (region 1) show that for both feature combinations the SVM performs better than CRF, respectively. But when using neighborhood information the CRF outperforms the SVM. Comparing the different feature combinations the prosodic feature combination (a) worked best for CRF, while the prosodic feature and RASTA-PLP combination worked best for SVM.

The best results for SVM over all speakers is 74.46% (84.26%), and for CRF 77.65% (86.27%). The best result for a single subject is 88.95% (92.86%), as shown in table 2. The RASTA-PLP improved the classification for the SVM, but for the CRF the detection rate was higher most times without them. Regarding the results for the combination of prosodic, RASTA and visual features averaged over all speakers, the combination with visual features were less good than without them. However, for one speaker adding the visual features improved the results, as can be seen in Table 2.

Table 2: *F1-score and unweighted accuracy for the single-, triple- and quintuple feature for classification with SVM and three different feature combinations for the best subject.*

region	1	4	7
SVM, [F1 (UA)]			
a	71.50 (85.75)	78.70 (87.60)	83.30 (90.40)
b	88.10 (91.60)	**88.90 (92.85)**	88.90 (92.80)
c	84.20 (89.45)	86.90 (91.95)	80.10 (84.75)
CRF, [F1 (UA)]			
a	57.64 (72.22)	87.92 (93.34)	86.77 (92.18)
b	79.07 (87.18)	88.68 (93.36)	87.43 (92.68)
c	77.19 (85.50)	**88.95 (92.86)**	88.21 (91.72)

6. Discussion

We first observe that for both methods, using context information - especially triple features and quintuple features - leads to a notable reduction of the detection error. That shows that the assumption of neighboring words changing by marking a word as prominent as described in [12, 13], can be found in our database as well and, therefore, should be used for classification. Considering only the fundamental frequency and the energy of the neighboring words already improves the results notably.

The experiments with CRF showed that labeling the neighboring words of prominent words as a separate category, leads also to an improvement, showing that there is not only a difference between prominent words and their surrounding but also between the direct neighboring words and the other not prominent words.

Out of both methods, the CRF performs best when incorporating context information. The different feature combinations show that RASTA-PLP features are useful for SVM but not for CRF in our experiment.

Using visual DCT features could only improve the performance for one single speaker. We suspect that this is due to the fact that the nose tracking, used to determine the mouth region, is not accurate enough and leads to errors. Overall, we see the best results for CRF using the feature combination (a) (only acoustic prosodic features without RASTA-PLP and DCT features) with the full context and trained on three label classes.

7. Conclusion

We tested different classification methods, SVM and CRF, on a database with 8 subjects and 3 different feature combinations in order to differentiate prominent and non-prominent words. Additionally we tested different amounts of context information.

For the CRF and SVM, considering the neighboring words demonstrated to lead to significant improvements. That confirms the hypothesis that prominence is not only determined by the word itself but also by the context in which it is embedded. By changing one word to be prominent, also the neighboring words are changed. The CRF performed better in our experiment than the SVM, which shows that the sequence information in a sentence (a hyperarticulated word is often embedded between two hypoarticulated words as well as it is not probable that a emphasized word is followed by another) contains much information. Using a separate label for the neighboring words of the prominent words even improves the performance.

Overall, the sequence based method CRF performs better than the SVM and the information in the neighboring words is beneficial for improving performance and should be utilized for the detection of prominent words. From the different feature combinations, prosodic features worked best for CRF and prosodic features with RASTA-PLP worked best for SVM. Averaged over all speakers the best results we get is an F1-score of 78.46%.

8. Acknowledgments

We thank Heiko Wersing and Petra Wagner for fruitful discussions. Furthermore, we are very grateful to Venkatesh Kulkarni for integrating the RASTA-PLP features into our framework. Many thanks to Mathias Franzius for support with tuning the SVMs and Merikan Koyun and Paschalis Mikias for the help with the data preparation.

9. References

[1] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," *European Conf. on Speech Communication and Technology (Eurospeech)*, pp. 1781–1784, 2005.

[2] M. Swerts, D. Litman, and J. Hirschberg, "Corrections in spoken dialogue systems." *INTERSPEECH*, pp. 615–618, 2000.

[3] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van, and E. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, pp. 339–373, 2000.

[4] J. Hirschberg, D. Litman, and M. Swerts, "Characterizing and predicting corrections in spoken dialogue systems," *Comput. Linguist*, vol. 32, pp. 417–438, 2006.

[5] J. Hirschberg, Litman, D., and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, vol. 43, no. 1-2, pp. 155–175, 2004.

[6] I. Bulyko, K. Kirchhoff, M. Ostendorf, and J. Goldberg, "Error-correction detection and response generation in a spoken dialogue system," *Speech Communication*, vol. 45, no. 3, pp. 271–288, 2005.

[7] G. Levow, "Identifying local corrections in human-computer dialogue," *In Eighth International Conference on Spoken Language Processing*, 2004.

[8] M. Heckmann, "Inter-speaker variability in audio-visual classification of word prominence," *INTERSPEECH, Lyon, France*, 2013.

[9] G. Levow, "Context in multi-lingual tone and pitch accent recognition," *INTERSPEECH*, pp. 1809–1812, 2005.

[10] M. Gregory and Y. Altun, "Using conditional random fields to predict pitch accents in conversational speech," *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pp. 677–683, 2004.

[11] G. Levow, "Automatic prosodic labeling with conditional random fields and rich acoustic features," *IJCNLP*, pp. 217–224, 2008.

[12] M. Dohen and H. Loevenbruck, "Pre-focal rephrasing, focal enhancement and post-focal deaccentuation in french," *Interspeech*, 2004.

[13] Y. Xu and C. Xu, "Phonetic realization of focus in english declarative intonation," *Journal of Phonetics 33*, pp. 159–197, 2005.

[14] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in *HLT-NAACL*, pp. 81–84.

[15] A. Rosenberg, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.

[16] M. Heckmann, "Audio-visual evaluation and detection of word prominence in a human-machine-interaction scenario," *INTERSPEECH, Portland, OR*, 2012.

[17] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[18] M. Heldner, "On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in swedish," *Journal of Phonetics*, vol. 31, no. 1, pp. 39 – 62, 2003.

[19] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans Speech and Audio Proc.*, vol. 2, no. 4, pp. 578–589, 1994.

[20] E. Fosler-Lussier, Y. He, P. Jyothi, and R. Prabhavalkar, "Conditional random fields in speech, audio, and language processing," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1054–1075, 2013.

[21] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[22] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, 2001.

[23] C. Sutton, "Grmm: Graphical models in mallet," http://mallet.cs.umass.edu/grmm/, 2006.