Automatic Prominence Classification in Swedish

Samer Al Moubayed, G. Ananthakrishnan, Laura Enflo

Centre for Speech Technology, Royal Institute of Technology (KTH),
Lindstedtsvägen 24, SE-10044, Stockholm,
{sameram, agopal, lenflo}@kth.se

Abstract

This study aims at automatically classifying levels of acoustic prominence on a dataset of 200 Swedish sentences of read speech by one male native speaker. Each word in the sentences was categorized by four speech experts into one of three groups depending on the level of prominence perceived. Six acoustic features at a syllable level and seven features at a word level were used. Two machine learning algorithms, namely Support Vector Machines (SVM) and memory based Learning (MBL) were trained to classify the sentences into their respective classes. The MBL gave an average word level accuracy of 69.08% and the SVM gave an average accuracy of 65.17 % on the test set. These values were comparable with the average accuracy of the human annotators with respect to the average annotations. In this study, word duration was found to be the most important feature required for classifying prominence in Swedish read speech.

Index Terms: Swedish prominence, SVM, MBL, syllable and word level features, word duration

1. Introduction

Prominence is a prosodic phenomenon which has been studied intensively in different languages. A linguistic segment is typically defined as prominent when it is perceived to stand out of its context [1]. Although prominence is qualitative according to this definition, many studies in different languages have targeted the quantification of prominence in terms of its production and perception of various prosodic, phonetic and linguistic correlates (Refer to [2] for a review on different languages).

The detection and quantification of prominence in speech plays an important role in many applications, since it concerns the question of *how* speech is produced and segments are contrasted, e.g. prominence aids the decoding in speech recognition, and hence can be used for syntactic parsing [3]. Recently, more research is focusing on the audio-visual relation of prosody. Many studies report findings on correlations between acoustic prominence and facial movements and gestures [4]. Hence, developing systems for automatic prominence estimation or classification is important for the use in speech recognition, speech synthesis, and audio visual speech synchrony systems.

In Swedish, prominence is often categorized with three terms: 'stressed', 'accented' and 'focused'. Previous research has reported that the most consistent acoustic correlate of stress in Swedish is segmental durations [5] and, less consistently, intensity [6]. As for accented syllables, the most apparent acoustic correlate compared to an unaccented foot is the presence of a fundamental frequency (F0) fall, referred to as a word accent fall [7]. Thus, an accent with a higher prominence level than 'stressed' is signaled mainly by F0, although an accented foot is usually also longer than an unaccented one [6]. Finally, in focal accent, which is generally considered the highest level of prominence, the primary acoustic correlate for distinguishing 'focused' from 'accented' words is a tonal one – a focal accent or a sentence accent rise following the word accent fall [7]. However, this F0 movement is usually accompanied by an increased duration and intensity of the word in focus [5, 8]. These studies suggest possible distinct acoustic realizations of prominence. Nonetheless, they distinguish different classes of prominence in terms of a structural model related to its underlying linguistic unit. For example, in 'focal accent', more than one syllable gets affected by different acoustic variations, which is specifically evident in compound words. In accented words, the acoustic realizations might or might not extend to more than one syllable.

This paper aims at automatically classifying levels of prominence in a speech segment, using sets of syllable and word level acoustic features, to train Support Vector Machines *(SVMs)* and Memory Based Learning *(MBL)*.

The paper is organized as follows: Section 2 presents the database and the annotation scheme; Sections 3 and 4 discuss the details of the method and features used for automatic classifications. Section 5 describes the experiments with the two machine learning algorithms used. In Section 6 we discuss the findings of the experiments and present the conclusions of the paper in Section 7.

2. Data and Annotation

Since it is suggested that prominence is perceived on a word level, while its acoustic correlates are realized at a syllable level, data collection for this problem has inherent limitations. Collecting annotations of prominence at a syllable level is firstly not relevant, since it disregards the perceptual prominence of the full word, and secondly, requires copious amounts of time for annotating large corpora. Acquiring annotations of prominence at a word level, introduces the risk of feature inconsistency at a syllable level, since in a prominent word, not all syllables are prominent. Nonetheless, it remains the consistent method since it relies on the human perception of the qualitative prominence of a word.

In this study, a dataset was selected from a corpus containing 5000 sentences of news texts and literature, read by a professional Swedish male actor. The corpus contains high-quality studio recordings for the purpose of speech synthesis voice creation. From this dataset, 200 sentences were randomly chosen and transcribed using a state-of-the-art speech aligner for Swedish [9]. The aligner gives a label file for each of the phonemes and words in the speech segment along with their

duration. The 200 sentences were then annotated according to the level of prominence perceived by four speech experts. The annotators used a visual tool to listen to the sentences and to mark each word as either *Prominent*, 'yes' (class label 2), *Not Prominent*, 'no' (class label 0), or *Maybe Prominent*, 'maybe' (class label 1). They were instructed to annotate prominence as a prosodic perceptual target, disregarding the underlying linguistic content. After collecting the annotations, the average answer (x) of the four subjects was considered as the prominence level of the word. This average is then rounded into three levels: 0 (No prominence) when $x<0.5$, 1 (Maybe prominent) when $0.5 \leq x \leq 1.5$, and 2 (Prominent) when $x>1.5$. In this approach, the class '*maybe'* is hypothesized to represent the guessing range of the subjects (when the subjects were not sure if a word is prominent or not, they chose the class *'maybe'* as the prominence level of this word). Table 1 presents the averaged confusion matrix between annotations of the different annotators against the average annotation.

The 200 sentences consisted of 2244 words and 3616 syllables. Out of the 200 sentences, 150 sentences were randomly chosen for training and cross-validation, and 50 for final testing.

Table 1*: The average confusion matrix between annotations of the different annotators against the average annotation*

Predicted Class	0 (no)	1 (maybe)	2 (yes)
True Class			
0 (no)	0.9744	0.3740	0.0399
1 (maybe)	0.0256	0.3669	0.2263
2 (yes)	0	0.2591	0.7337

3. Method

The machine learning algorithms were trained to classify prominence using syllable and/or word level features. In the case that syllables were classified for prominence, the classification results of all the syllables of a word were integrated to give the final prominence at a word level. All the syllables of a word carry their word's prominence during training. This strategy could be called a late integration approach. On the other hand, in an early integration approach, syllable level features were combined to give word level features, before the training of the machine learning algorithms.

Two major learning algorithms were used, namely Support Vector Machines, and Memory Based Learning, which represent examples of what is usually referred to as 'eager learning' and 'lazy learning', respectively.

3.1. Post-Integration of Prominence from Syllable to Word level

Among the various late integration methods, four different methods were tried, as follows:

Average: The mean value for all of the classes is rounded to its nearest integer.

WAverage: Presence of prominence is given more weighting (i.e. yes, maybe and no are given weighting 3, 2 and 1, respectively).

Max: The maximum value from the entire word is picked. If one syllable in the word is prominent, the word is considered prominent.

Voting: A non-democratic voting scheme, where prominent syllables get more voting rights to choose the classification of the word (every "yes" has 3 votes, "maybe" has 2 votes and "no" has 1 vote).

4. Feature Set

As mentioned in the introduction, duration, loudness and F0 movements are major acoustic correlates of prominence. In this study, a wide set of both syllable level and word level features have been taken into account. Features representing the syllables are taken from the syllable vowel, since vowels represent the nuclei and the acoustically stable part of the syllable:

- ***Vowel Type***
- ***Vowel Duration***
- ***Mean F0:*** F0 is calculated for all the vowels in the files using the YIN state of the art real-time F0 tracking algorithm [10].
- ***Vowel Mean Delta F0:*** Since F0 movements have been shown to play a major role in realizing prominence, the mean absolute value of the log delta F0 is calculated for each of the vowels.
- ***Vowel Loudness:*** The study in [11] on British English reports that if a reliable measure of loudness is used, loudness can become a significant parameter to estimate prominence. In [12], it is found that spectral emphasis in Swedish correlates to prominence. In this study we estimate the vowel average loudness using the ITU [13].
- ***Vowel Spectral Tilt:*** The role of spectral tilt as a prominence correlate in English has been investigated [14], as it is considered a good representative of the spectral quality of speech. Hence, the spectral tilt was calculated over the vowels with a frame centered at the middle of the vowel.
- ***Average Syllable Duration In the Word***
- ***Word Duration***
- ***Number of syllables***
- ***Word Mean F0***
- ***Word Mean Delta F0***
- ***Word Mean Spectral Tilt***
- ***Word Loudness***

5. Experiments with SVM and MBL

5.1 Baseline

Although it was expected that a very low percentage of words would be marked prominent, almost half the database consisted of words which were either 'maybe' prominent or prominent ('yes'). We found that 56.27% of the words in the validation set and 47.36% of the words in the test set had 'no' prominence, which was used as the baseline to compare the results of the classification task.

5.2 Memory Based Learning (MBL)

In this study, the Tilburg Memory-Based Learner (TiMBL) [15], which is a decision-tree-based implementation of k-nearest neighbor classification, was used. Very few abstractions are formed using the training data (lazy learning), but local

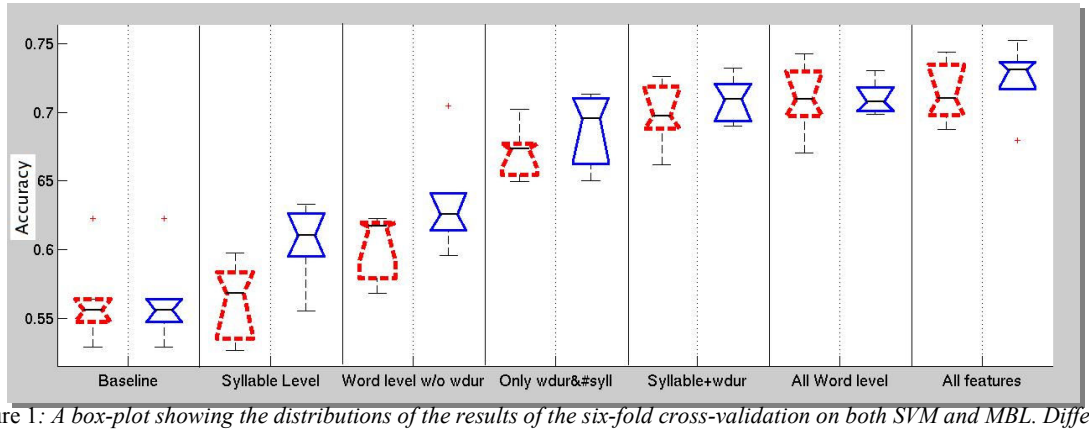

Figure 1: *A box-plot showing the distributions of the results of the six-fold cross-validation on both SVM and MBL. Different combinations of the features are compared to the baseline. Red: MBL, Blue: SVM.*

generalizations are formed for a particular test sample, in order to classify it. For the numeric features Euclidian distance metric was used and for the symbolic features the Modified Value Distance Metric (MVDM) [16] was applied.

5.3 Support Vector Machines

Support Vector Machine (SVM) is a commonly used 'eager learning' method. It finds the best separating hyper-plane between two sets of classes in such a way that the distance between the two classes is maximized. Using different kinds of kernel functions, the separating hyper-plane can be found in a space of higher dimensionality than the data itself. It performs especially well with sparse data.

In this study, we used an implementation of Least Square SVM (LS-SVM) [17]. For features which were symbolic, MVDM bins were pre-calculated to make them numeric values with one feature for each class. The features were then normalized for mean and standard deviation before training the SVMs. The results on the validation set are shown in Table 4 for the optimized set of parameters.

6. Analysis and Results

By running a six-fold cross validation on the training corpus for both SVM and MBL, the classification accuracy for single features are presented in Table 2. It is clear that there is a high correlation between the word duration, the number of syllables in the word and the prominence of the word. The SVM gave a baseline performance whenever the features could not contribute to the classification.

The SVM and MBL methods were also trained on other combinations of the features, most importantly: 1-Only syllable level features with post integration; 2- Only word level features; 3- Syllable level and word level features with post integration. Figure 1 presents a box-plot showing the distributions of the accuracies of the six-fold cross-validation for different combinations of the features, and Table 3 presents the mean values of the accuracy on the same combinations for both MBL and SVM.

Among the different late integration schemes suggested, we found that none of the schemes performed consistently better. Hence, we have presented the average results from the different integrations schemes suggested. Another important aspect of each of the methods is the confusion matrix, which indicates the per-class accuracies as opposed to the overall accuracy for the three prominence classes. Table 4 and Table 5 display the confusion matrices for both the MBL and the SVM for the best feature set (all features); with an overall accuracy on the 50 sentences in the test set is 69.08% for MBL and 65.17% for SVM.

Table 2: *Table showing the classification accuracy on single features using SVM or MBL sorted by their accuracies*

Feature	MBL Accuracy	SVM Accuracy
Word Duration	64.58	66.65
No Syllables	61.69	62.25
Average Syllable duration	56.86	59.04
Vowel Type	55.84	58.90
Word Mean Delta F0	54.55	52.67
Vowel Duration	54.10	52.67
Word Spectral Tilt	52.15	52.67
Word Loudness	48.29	52.67
Word Mean F0	47.99	52.67
Vowel Mean Delta F0	46.43	52.67
Vowel Average F0	44.97	52.67
Vowel Spectral Tilt	42.79	52.67
Vowel Loudness	41.01	52.67

Table 3: Average *accuracies for different feature combinations on the six-fold cross-validation for the two methods (Baseline word accuracy = 52.67)*

Features	MBL (%)	SVM (%)
Syllable Features	56.86	61.24
Word Features	70.98	71.03
Syllable Features + Word duration and No. Syllables	70.12	71.23
Only Word Duration and No. Syllables	67.18	68.77
All Features	**71.52**	**72.55**

Table 5: *Confusion matrix for the best features using the MBL on the test data*

Predicted Class	0 (no)	1 (maybe)	2 (yes)
True Class			
0 (no)	85.12	7.85	7.02
1 (maybe)	35.95	25.49	38.56
2 (yes)	9.48	14.66	75.86

Table 6: *Confusion matrix for the best features using the SVM on the test data*

Predicted Class	0 (no)	1 (maybe)	2 (yes)
True Class			
0 (no)	87.19	8.26	4.55
1 (maybe)	31.37	41.18	27.45
2 (yes)	7.76	24.14	68.10

7. Discussion

Only 40% of the syllables had 'no' prominence, while classes 'maybe' and 'yes' had around 30% syllables each. At the word level, 47% of the words had no prominence, 22% of the words had 'maybe' prominence and 18% of the words had 'yes' prominence. This indicated that words with a higher number of syllables tended to be more prominent in the Swedish language. Therefore, the number of syllables in a word was an important feature for the classification task. This indication was vindicated by the fact that the most informative features were word duration and number of syllables in the word as shown in Table 2 and Figure 1. Most of the syllable level features did not perform better than the baseline, while word duration performed considerably better and contributed largely to the final accuracy.

As shown in Figure 1, classification with only syllable level features using SVM was significantly better than the baseline ($p=0.03$), but no improvement was observed for MBL. Using word level features without word duration and number of syllables was almost the same as using only syllable level features for SVM, but was significantly better for MBL ($p=0.02$). The addition of either syllable level features or other word level features to word duration boosted the performance by a small amount. However, using all the features boosted the performance significantly over using only word duration ($p=0.03$ and 0.004 for SVM and MBL respectively).

Even though the final results showed MBL in a slightly better light, the variations in MBL for different parameters of the algorithm were large, showing that the optimization was highly crucial and difficult. On the other hand, SVM seemed a more convenient tool to use because of the considerably fewer number of parameters to tune and the stable performance for different parameters.

The results in general also show that the confusions were higher for the 'maybe' class, while the confusions between 'no' and 'yes' classes were not as high (2-6%). This gives a possibility for higher accuracies in automatic prominence detection using only two levels. Looking at Table 1, which is the confusion matrix of the annotators with respect to the average annotations, we can see that there is high confusion in the 'maybe' class, as expected. In addition, it is also interesting to see that these properties are similar to what we observe in the confusion matrices of the two machine learning algorithms.

8. Conclusion

Four annotators have categorized a Swedish database of read speech, consisting of 200 sentences by one male speaker, into three levels of prominence. Their agreement with the average annotations of prominence had an average of 69.17%. Using six acoustic features at a syllable level and seven acoustic features at a word level, we trained two machine learning algorithms to automatically classify the prominence of the words in the sentences. The Memory Based Learning and Support Vector Machines gave an average accuracy of 69.08 and 65.17%, respectively, on the test data. It was found that the word duration was the most significant feature for this classification task. The other acoustic features at a syllable level and word level did not contribute significantly to the results when used in isolation, but boosted the performance when used along with word duration. The final performances of the machine learning algorithms were comparable to the average agreement of the human annotators. Thus it may be possible to automatically annotate databases of Swedish read speech for prominence levels.

9. Acknowledgements

Part of this work has been supported by the Graduate School of Language Technology (GSLT), Sweden. We would also like to thank Walter Daelemans for providing help on TiMBL and Joakim Nivre for his guidance.

10. References

[1] Terken, J., "Fundamental frequency and perceived prominence of accented syllables". The Journal of the Acoustical Society of America, 1991.

[2] Horne, M., "Prosody: "Theory and Experiment: Studies presented to Gösta Bruce", Springer, 2000 (ISBN: 978-0-7923-6579-2).

[3] Wang, D., Narayanan. S., "An acoustic measure for word prominence in spontaneous speech". IEEE Transactions on Audio Speech and Language Processing, 2007.

[4] Al Moubayed, S., & Beskow, J., "Effects of Visual Prominence Cues on Speech Intelligibility", In Proceedings of Auditory-Visual Speech Processing AVSP'09, Norwich, England, 2009.

[5] Fant, G., and Kruckenberg, A. "Notes on stress and word accent in Swedish", STL-QPSR, 35(2-3), 125-144., 1994.

[6] G. Bruce, Swedish word accents in sentence perspective. LiberLäromedel/Gleerup, 1977.

[7] Bruce, G. and Hermans, B. "Word tone in Germanic languages" . Empirical Approaches to Language Topology. p 605—658. 1999

[8] Heldner, M., and Strangert, E., "Temporal effects of focus in Swedish" Journal of Phonetics, 29(3): 329-361, 2001.

[9] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech," in Proceedings of Fonetik. 2003.

[10] A. de Cheveigne and H. Kawahara. "YIN, a fundamental frequency estimator for speech and music", The Journal of the Acoustical Society of America, 111:1917, 2002.

[11] Kochanski, G. and Grabe, E. and Coleman, J. and Rosner, B. " Loudness predicts prominence: Fundamental Frequency lends little", The Journal of Acoustical Society of America, v118, p1038. 2005.

[12] Heldner, M. "Spectral emphasis as an additional source of information in accent detection", Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding, NJ, 57-60, 2001.

[13] ITU-R BS.1770-1 Algorithms to measure audio programme loudness and true-peak audio level (2006-2007)

[14] Campbell, N. and M. Beckman: Stress, prominence and spectral tilt. In A. Botinis, G. Kouroupetroglou and G. Carayannis (eds), Proceedings of the ESCA Workshop on Intonation: Theory, Models and Implications, Athens, Greece, September 18-20, 1997, p. 67−70. Athens: ESCA and University of Athens. 1997.

[15] Daelemans, W. and Zavrel, J. and van der Sloot, K. and van den Bosch, A., "Timbl: Tilburg memory-based learner" version 6.2.0, 2007, http://ilk.uvt.nl/software/

[16] Mäntaras, R.L., "A distance-based attribute selection measure for decision tree induction", Machine learning, 6(1): 81-92, 1991.

[17] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, 'Least Squares Support Vector Machines', World Scientific Pub. Co., Singapore, 2002 (ISBN 981-238-151-1).