



Prominence Detection in Swedish Using Syllable Correlates

Samer Al Moubayed and Jonas Beskow

KTH, Centre for Speech Technology CTT, Stockholm, Sweden

sameram@kth.se, beskow@kth.se

Abstract

This paper presents an approach to estimating word level prominence in Swedish using syllable level features. The paper discusses the mismatch problem of annotations between word level perceptual prominence and its acoustic correlates, context, and data scarcity. 200 sentences are annotated by 4 speech experts with prominence on 3 levels. A linear model for feature extraction is proposed on a syllable level features, and weights of these features are optimized to match word level annotations. We show that using syllable level features and estimating weights for the acoustic correlates to minimize the word level estimation error gives better detection accuracy compared to word level features, and that both features exceed the baseline accuracy.

Index Terms: prominence, accent, focus, syllable emphasis

1. Introduction

Prominence is traditionally defined as when a segment (a syllable, a word, or a phrase) stands out of its context, as defined by [1]. Others use it as the perceptual salience of a linguistic unit [2]. There have been numerous studies on the perception, production and modeling of prominence and its phonetic and prosodic correlates in different languages.

The detection and quantification of prominence phenomena in speech can play an important role in many applications since it concerns the question of how speech is produced and how segments are contrasted, e.g. decoding in speech recognition, and hence can be used for syntactic parsing [3], speech comprehension [4]. More recently, research is focusing on the audio-visual function of prosody, for example, in [5] it was found that visual prominence can enhance speech comprehension if coupled with the acoustic. Hence, it is important for speech driven avatars to detect prominent segments in the speech signal to drive gestures.

Research has investigated the acoustic-prosodic cues to prominence on a syllable or on a word level, some using lexical and higher level linguistic information. This study focuses on estimating prominence using syllable based segments, since in some applications, the word level segmental information might not be available. This presents a theoretical challenge since such a method requires sufficient information about prominence inside the boundaries of the phonetic segment. In addition, some prominence categories (levels) are perceptually based on a word level, and hence reliably transcribed data on a syllable or vowel level is not available.

The paper is organized as follows: Section 2 presents an overview on the Swedish acoustic prominence model in the literature. Section 3 presents the data collection and annotation. Section 4 presents the proposed prominence model. Section 5 discusses prominence feature extraction. Section 6 presents experiments and results on using the proposed model, and section

7 discusses the results and concludes the paper.

2. Acoustic Correlates to Prominence in Swedish

In Swedish, prominence is often categorized with three terms: *stressed*, *accented* and *focused*. Previous research reported that the most consistent acoustic correlate of stress in Swedish is segmental durations [6]. In addition, overall intensity differences have also been studied among the correlates of stress [6]; although these differences may not be as consistent as the durational differences [7]. As to Accented syllables, according to the Swedish intonation model, the most apparent acoustic correlate for accented from an unaccented foot is the presence of an f0 fall, referred to as a word accent fall. Thus, accent as a higher prominence level than just stress is signaled mainly by f0, although an accented foot is usually also longer than an unaccented one [6]. Finally, in focal accent, which is the highest level of prominence, the primary acoustic correlates for distinguishing *focused* from *accented* words is a tonal one (a focal accent or a sentence accent rise following the word accent fall [8]). However, this f0 movement is usually accompanied by an increased duration of the word in focus [9], and by moderate increases in overall intensity [10]. Hence, according to the Swedish intonation model, f0 movements should be considered in any prominence modeling system, but they are by no means the only existing correlates.

3. Data Collection

In Swedish, as already mentioned, the production of prominence (and hence the perception too) can happen on a syllable or on a word level, and that is mainly decided by the level and the type of prominence. For example, focal accent on a compound polysyllabic word is realized acoustically over several syllables (mainly the first and the last ones). This introduces several difficulties, firstly that word boundaries are needed for any method to accurately detect the correct type and level of prominence over that word; secondly that if annotations of prominence are collected over a word level, this will disregard the details about on which syllable(s) the acoustic correlates to stress are realized, on the other hand, collecting annotations on a syllable level is slow and more time consuming, and disregards the word level perception of prominence.

In this work, we undertake the approach of collecting word prominence rather than syllable prominence, this is mainly done due to that it is faster, easier and practical for annotators, and it captures the perception of prominence when produced over words (lexical items) rather than over syllables.

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 phonetically aligned. 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 clicked on each word to increase its prominence rating. The buttons were presented in white color stating that they are "Not Prominent". The annotators were instructed to click one time on the word if they thought that it is "maybe prominent", and the button turned grey, or to click twice to indicate that the word is "Prominent" and the button turned green. They were also instructed to annotate prominence as a prosodic perceptual target, disregarding the underlying linguistic content. Figure 1 presents a screenshot of the tool used to collect the annotations. This provided annotations per word on three levels, which were then transformed numerically to (0, 0.5, 1) according to (prominent, maybe prominent, not prominent) accordingly. After collecting the annotations, an averaging approach was used by giving each word the average of the four annotators (this approach gives equal weight to all the annotators, unlike other approaches like majority voting).

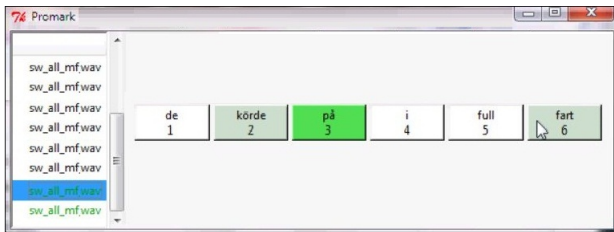


Figure 1: A screenshot of the visual annotation tool used to annotate prominence, where every word is associated with one button.

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. Figure 2 shows the histogram of the average prominence value over annotators and words.

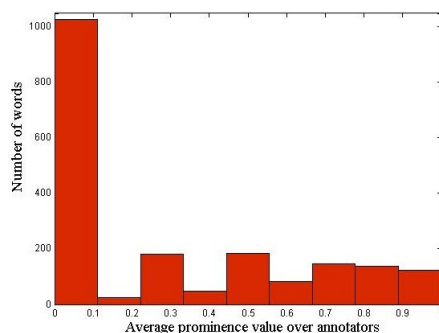


Figure 2: The histogram of the prominence value per word after averaging over the 4 annotators.

4. Prominence Model

In the last couple of decades, since collecting large amounts of data became practical, research on modeling prosodic events has employed data-driven approaches where non-linear models

Table 1: The averaged confusion matrix between annotations of the different annotators (columns) against the average annotation (rows).

-	Not Prom.	Maybe Prom.	Prom.
Not Prom.	0.9744	0.3740	0.0399
Maybe Prom.	0.0256	0.3669	0.2263
Prom.	0	0.2591	0.7337

are fit to the data. These models have been used in several experiments on detecting pitch accents in speech, as using Gaussian Mixture Models [11], and Neural Networks [2].

Nonetheless, these approaches require intensive amounts of annotated corpora on prominence; moreover, as discussed earlier, in Swedish, the prosodic prominence structure is realized on syllable level while the perception is realized on a word level, this means that if features are extracted on a syllable level, they will lack a syllable level annotation, and using word level features is firstly context demanding (where detection of word boundaries should be done on the acoustic signal), and secondly, word level features will introduce inconsistency since prominence is not realized over the whole word but on specific syllables in it.

In this work, we suggest a linear model for modeling syllable prominence, and post integration of the estimated syllables prominence will estimate the underlying word prominence. Linear simple models have already shown good performance, as in [12] on English read the spontaneous speech.

If x is an observation sample from a random parameter X , we define the level of prominence of x as:

$$Prom_x = 1 - f_X(x), f \in [0..1] \quad (1)$$

Where f_x is the normalized likelihood of x in the random distribution of X . If x is a feature vector of n independent features $x = \{x_1, x_2, \dots, x_n\}$, the prominence level of the observation vector x is:

$$Prom_x = 1/n \sum_{i=1}^n w_{x_i} Prom_{x_i} \quad (2)$$

While w_{x_i} is the weighting of x_i .

Since the data has only word level annotations, in the case when x is a syllable level observation, we assume that the syllable with the higher prominence level in the word containing it provides the prominence of the word. In this case, the weights w_{x_i} can be tuned to minimize the error of estimating the word prominence when using only syllable level features.

It is good to note that this definition assumes prominence on a continuous scale without any structural boundaries between different types of prominence. This view of prominence has been suggested and used previously, as in [13].

5. Feature Set

As shown in the Swedish prominence model above, duration, loudness and F0 movements are major acoustic correlates of prominence. In this study, sets of either syllable level or word level features have been used. Features representing the syllables are taken from the syllable vowel, since vowels represent the nuclei and the acoustically stable part of the syllable. The data used in this study is taken from one male speaker thus no subject specific feature normalization is applied.

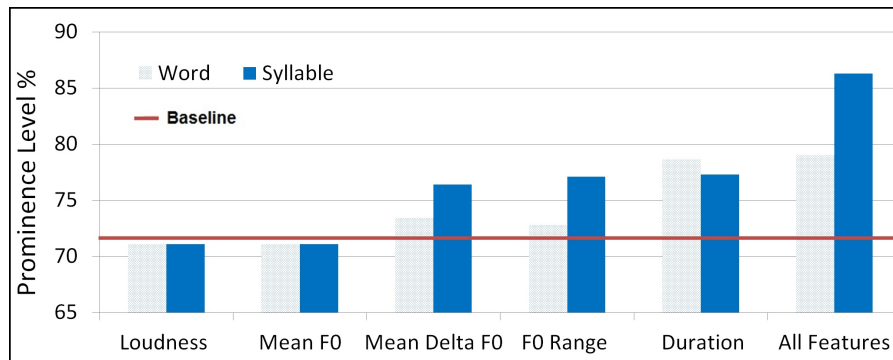


Figure 3: The prominence detection percentage accuracies for single and multiple features against the baseline. The baseline gives a fixed value of the average prominence of all the words in the dataset

5.1. Fundamental Frequency

Fundamental frequency movements are important parameters in the perception of prominence, and pitch accents have been under interest in many prominence modeling and detection research in many languages [14]. In Swedish, accent fall and rise characterize *accented* and *focused* segments. Moreover, in a study by Heldner [9], it was found that the faster and the higher the accent rise, the more agreement among annotators on the prominence level of the segment. This entitles a vowel based prominence detection system to capture the movements of the fundamental frequency on the segment. For this we use three F0 features: *Average F0*, *Average delta F0* which is the averaged first derivative of F0 over the vowel, and *F0 range* over the vowel.

5.2. Duration

The segmental duration, where in this study, the duration of the word or the vowel is an acoustic feature.

5.3. Loudness

In this study we estimate loudness using the *ITU-R1077* [15] recommended loudness measure, which was recently evaluated to be a highly reliable perceptual measure by [16], this measure applies a high band and low band frequency filters and is computationally very efficient. In this method, this measure is adapted to calculate the mean loudness over a window sized as the length of the underlying segment, be it a vowel or a word.

6. Experiments and Results

The speech files from the 200 sentences dataset were phonetically aligned, and the acoustic features, namely: duration, loudness, average F0, average delta F0 and F0 range were extracted over all the syllables and over all the words. Half Gaussian distributions were estimated for each of the features, and the likelihood of each feature was calculated, and hence the prominence level per features. The data in all the tests was split into 4 cross validation folds giving 25% test set size. The baseline in this study was simply taken as the average word prominence of the whole dataset. This gives a fixed prominence value for all the words in the test set. For this dataset the average baseline value was 0.3, which means that the average word prominence in the database was 0.3.

6.1. Word Features

For words, weights were estimated using a grid-search method which searches in the range [0..1] for an optimized weight for each feature so that the average weighted prominence for each word is optimized. Since the features are on a word level, and word level annotations are available, only one weight per feature was optimized.

6.2. Syllable Features

As for syllables, the weights were estimated by assuming that the syllable which has the highest averaged weighted prominence holds the underlying word prominence, in this case, the weights were optimized using a grid-search optimization to minimize the error on this post-integration function. An illustration on the estimation of prominence using syllable features is presented in Figure 4. In the figure, the prominence of each word is taken as the highest prominence of all its syllables.

6.3. results

The 100% percentage accuracies for estimating word prominence for single features and for syllable and word features are presented in Figure 3,

By looking at the results, it is shown that several single features are unable to increase the accuracies beyond the baseline; it is good to mention that, as shown before, the distribution of the prominence values per word is highly unbalanced where most words are not prominent, and this simple baseline gave an accuracy of 72%. Nonetheless, syllable F0 range performed singly better than all word features combined and the baseline, and the performance was increased for the syllable duration. The best performance of 86% was reached using all the syllable features combined, with optimized weights equal to: Loudness: 0.13, Duration: 0.52, delta F0: 0.21, F0: 0, and F0 range: 0.14.

7. Discussion and Conclusions

In [17], a comparison on English between detecting prominence on a vowel, syllable, and word level shows that given more context, the information available to detecting prominence is increased, and the estimated model is fine-tuned the more segmental information is given. This increases the need for word level segmental information in a language like Swedish, where perceptually it was shown that there are more than changing segment parameters to prominence, but prominence follows a

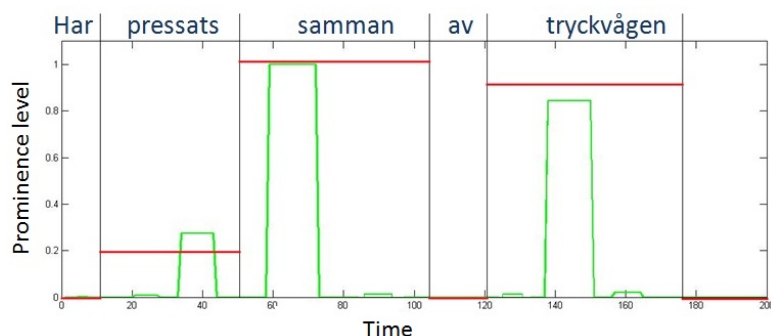


Figure 4: An example of an output sentence. Red lines represent the word level original annotation. Green lines represent the automatic estimation of prominence per syllable. "Has been compressed by the pressure wave"

structural and segmental model defined over time.

This work presents the first steps to build a syllable prominence detection method using a linear model of the acoustic correlates. In the experiments, it is shown that this model performs better when using syllable correlates and post integration of syllables prominence to estimate word prominence, than when word correlates are used in the same method. This might indicate that acoustical correlates of prominence on syllables are stronger than those on words. In [18], it was shown that, for prominence classification using Support Vector Machines (SVMs), syllable features can not contribute to accuracy beyond the baseline, while word lexical features like word duration, the number of syllables per word, can enhance the classification significantly; nonetheless, these features are lexical, and require word boundaries. The advantage of only using syllable (in this study vowel) features allows a system which only has phonetic boundaries and the speech segments to estimate prominence, without the need for any higher lexical or linguistic knowledge, as a case where prominence can be used to aid speech recognition (for example in word n-grams, or syntactic parsing). Although the proposed method is evaluated using one-speaker read speech with enough data to estimate the prosodic parameters distributions, it is interesting to extend the estimation of the model to a one in real-time, where the parameters distributions are calculated over a moving window.

8. Acknowledgements

This research is carried out at KTH Speech Music & Hearing and the Centre for Speech Technology, supported by Mon-AMI, an Integrated Project under the European Commissions 6th Framework Program (IP-035147).

9. References

- [1] J. Terken and D. Hermes, "The perception of prosodic prominence," in *Prosody: Theory and Experiment. Studies Presented to Gösta Bruce*, 2000, pp. 89–127.
- [2] B. Streefkerk, L. Pols, and L. Bosch, "Acoustical features as predictors for prominence in read aloud dutch sentences used in ann's," in *Sixth European Conference on Speech Communication and Technology*. Citeseer, 1999.
- [3] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 690–701, 2007.
- [4] M. Grice and M. Savino, "Can pitch accent type convey information status in yes-no questions," in *Proceeding of the Workshop Sponsored by the Association for Computational Linguistics*, 1997, pp. 29–38.
- [5] S. Al Moubayed and J. Beskow, "Effects of visual prominence cues on speech intelligibility," *Proceedings of the International Conference on Auditory Visual Speech Processing AVSP'09*, vol. 15, p. 16.
- [6] G. Bruce, B. Granstrom, K. Gustafson, M. Horne, D. House, and P. Touati, "On the analysis of prosody in interaction," *Computing Prosody: Computational Models for Processing Spontaneous Speech*, pp. 42–59, 1997.
- [7] G. Fant and A. Kruckenberg, "Notes on stress and word accent in swedish," in *Proceedings of the International Symposium on Prosody, Sept 18 1994, Yokohama*, 1994, pp. 2–3.
- [8] G. Bruce, *Swedish word accents in sentence perspective*. LiberLäromedel/Gleerup, 1977.
- [9] M. Heldner and E. Strangert, "Temporal effects of focus in swedish," *Journal of Phonetics*, vol. 29, no. 3, pp. 329–361, 2001.
- [10] G. Fant, A. Kruckenberg, J. Liljencrants, and S. Hertegård, "Acoustic phonetic studies of prominence in swedish," *KTH TMH-QPSR*, vol. 2, no. 3, 2000.
- [11] N. Obin, X. Rodet, and A. Lacheret-Dujour, "Prominence model: a probabilistic framework," in *submitted to The 33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP08), Las Vegas, USA*, 2008.
- [12] F. Tamburini, "Prosodic prominence detection in speech," in *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications*, vol. 1, 2003.
- [13] G. Fant, A. Kruckenberg, and J. Liljencrants, "The source-filter frame of prominence," *Phonetica*, vol. 57, no. 2-4, pp. 113–127, 2000.
- [14] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *The Journal of the Acoustical Society of America*, vol. 89, p. 1768, 1991.
- [15] "Recommendation itu-r bs.1770-1. algorithms to measure audio programme loudness and true-peak audio level."
- [16] P. Nygren, "Achieving equal loudness between audio files - evaluation and improvements of loud-ness algorithms," *Master's thesis, KTH CSC TMH*, 2009.
- [17] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 81–84.
- [18] S. Al Moubayed, G. Ananthakrishnan, and L. Enflo, "Automatic prominence classification in swedish," *Proceedings of Prosodic Prominence: Perceptual and Automatic Identification Workshop*, Chicago, U.S.A. 2010.