

An unsupervised approach to prominence estimation in Swedish

Samer Al Moubayed

KTH Centre for Speech Technology, Stockholm, Sweden.

sameram@kth.se

(Submitted as a term paper to the GSLT Statistical Learning Course.)

Abstract

This paper presents an unsupervised approach to detecting prominence in Swedish on a continuous scale using vowel acoustic features over time. By modeling vowel duration, mean loudness and mean delta pitch over the vowel, a prominence level for each vowel is estimated using a function of the likelihood of these parameters. An evaluation of the method to estimate word prominence results in 72.4% mean square error on read speech annotated by 9 speech experts, and 81.7% accuracy on focal accent detection on a small test set, given the number of focally accented words in the input utterance.

1. Introduction

Prominence is traditionally defined as when a segment (a syllable or a word) stands out of its context, as defined by Terken (1991), others use it as the perceptual salience of a linguistic unit (Streefkerk 1999). There have been numerous studies on modeling and perception 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 (Wang & Narayanan (2007)). 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. In Krahmer & Swerts (2007), it was found that visualizing manual gestures on acoustically prominent words increases the acoustic perception of prominence these words receives, moreover, producing visual prominence gestures correlates with an increased acoustic prominence of this word. Hence, it is important for speech-to-visual talking agents to detect prominent segments in the speech signal to drive gestures.

Many studies have investigated the acoustic-prosodic cues to prominence on a syllable or on a word level, and some using lexical and higher level linguistic information. In some situations, contextual information might not

be available. We are interested in this study to detect prominence using phone based segments, since in some applications, the segmental information about syllables and words might not be available (i.e. context independent phoneme recognizers). 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 word level, and hence reliably transcribed data on a syllable or vowel level are not available.

In this paper we present an unsupervised method to detecting the level of prominence on a vowel basis in Swedish, where vowel segments are recognized using a phoneme recognizer.

The rest of the paper is organized as follows: Section two gives an overview of the Swedish acoustic prominence model. Section 3 presents the proposed prominence function. Section 4 presents the proposed modeling of the acoustic parameters. Section 5 presents results from two experiments on detecting prominence, one using a mean square error of the estimated prominence over a word level prominence annotations, and the second evaluates the function on a binary classification task. Section 6 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 (Fant & Kruckenberg (1994); Bruce & Granström (1997)). In addition, overall intensity differences have also been studied among the correlates of stress (Bruce, 1999); although these differences may not be as consistent as the durational differences (Fant & Kruckenberg (1994)). 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 (Bruce, 1999). 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 (Bruce, 1977). However, this f0 movement is usually accompanied by an increased duration of the word in focus (Fant & Kruckenberg (1994), 1994;Heldner & Strangert (2001)), and by moderate increases in overall intensity (Fant et al. (2000)).

Hence, according to the Swedish intonation model, f0 movements should be considered as an important type of acoustic correlate of prominence, but they are by no means the only existing correlates.

From this background discussion, it seems reasonable that a prosodic prominence level detection system in Swedish should include acoustic features about segment duration, loudness, and an estimation of f0 movements. It is also important to note, that although “stress” is a syllable based event, accents are usually word based perceptual categories, and the acoustic correlates can be distributed over more than one syllable in poly-syllabic words (a word accent fall, followed by a final accent rise), this model makes it difficult to detect focal accent using only syllable or vowel information, but still, the correlates of a focal accent – e.g. longer syllable duration - might propagate outside the accented syllables, and exhibit changes in some, or all of the syllables of the word, and hence, using only vowel acoustic information might still be a feasible approximation.

3. Method for Modeling Vowel Prominence

We project the definition of prominence on acoustic parameters to hypothesize that a prominent observation of a parameter is an observation which exhibits an unexpected value. Statistically this would mean that the lower the likelihood of the value, the higher its prominence.

Let us consider, an observation x , a sample from a random parameter X , we define the prominence level of the observation x simply as:

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

Where $f(x)$ is a function of the likelihood of the observation 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 x becomes:

$$\Pr_x = \frac{1}{n} \sum_{i=1}^n \Pr_{x_i} \quad (2)$$

The assumption that the relation between the parameters’ prominence is simply additive is motivated since,

as described in the prominence model above, in general, none of the acoustic correlated to prominence (duration, loudness, f0 movements) is mandatory, and the higher prominence level seems to correlate with higher variation in one or more of these parameters.

This definition is still too general, as a general definition of prominence, since it might include other types of phenomena in speech like increased speech rate, hesitations, etc, that is when the parameters exhibit unusual values, and hence the choice of the parameters and their functions should reflect the constraints of the language specific prominence model.

On the other hand, this definition breaks the boundaries between the prominence categories, and estimates segment prominence on a continuous scale.

4. Feature Selection

This study aims at estimating prominence over vowels, since vowels represent syllables nuclei, and hence the acoustic feature vector is extracted over the vowel segment. It is worth to mention here that these parameters may be modeled over vowels as one category, or for each type of vowel which would result in a finer model compensating for the intrinsic differences in these parameters among vowels (i.e. intrinsic vowel duration, intrinsic vowel intensity), on the other hand, this requires larger amount of data for a reliable estimation of the model, which, for example, might be impractical in on-line modeling.

Following is a detailed description of the included acoustic features and their estimated distributions.

Duration

The increase in segmental duration or, as in this study, the vowel duration, seems to be a major correlate to prominence, be it stressed or accented. Since only the increase, but not the decrease, in vowel duration correlates to prominence, we model vowel duration as a random variable generated from a half-Gaussian distribution $(\mu_{dur}, \sigma_{dur}^2)$ with μ is the mean vowel duration and σ^2 the standard deviation.

By looking at the histogram of vowel duration in Figure 1, it is reasonable to assume that it is best fitted using a half-gaussian distribution.

Hence the prominence function defined in equation (1) for loudness becomes:

$$\Pr_x = 1 - \frac{L(x | \mu_{dur}, \sigma_{dur}^2)}{L(\mu_{dur} | \mu_{dur}, \sigma_{dur}^2)} \quad (3)$$

Where L is the half-Gaussian likelihood function, and \Pr_x is the complement of the normalized likelihood.

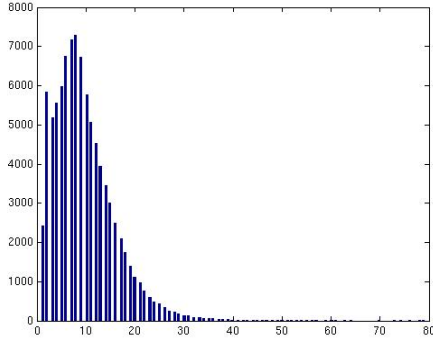


Figure 1. Histogram of vowel duration using 100,000 vowels.

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 (Terken (1991)). In Swedish, accent fall and rise characterize “accented” and “focused” segments. Moreover, in a study by Heldner (1998), it was found that the bigger the size of the accent rise, the more agreement among annotators on the prominence level of the word. This entitles a vowel based prominence detection system to capture the movements’ shape of the fundamental frequency on the vowel segment. From this we propose to represent the fundamental frequency prominence level over vowels using the mean delta pitch (MDP) calculated over the vowel segment. This is calculated for a vowel v of n successive F0 values as:

$$MDP_v = \frac{1}{n} \left| \sum_{i=2}^n \log(f0_v(i)) - \log(f0_v(i-1)) \right| \quad (4)$$

This function models the absolute value the speed of the pitch over a vowel but it is limited to capturing a unidirectional pitch accent, - i.e. symmetric movements of pitch are considered less prominent- and does not distinguish in the prominence level between a falling or rising accent. Figure shows the histogram of MDP using 100,000 samples.

We model this pitch feature as a half-Gaussian distribution (as we modeled loudness), with $\mu_{mdp}=0$ (flat pitch). The pitch prominence level function becomes:

$$\Pr_x = 1 - \frac{L(x|0, \sigma_{mdp}^2)}{L(0|0, \sigma_{mdp}^2)} \quad (5)$$

Loudness

As mentioned before, studies reported a possible correlation between intensity and prominence. In the literature, many studies have investigated the role of intensity measures and prominence, in Kochanski et al (2005), a study on British English reports that if a reliable measure

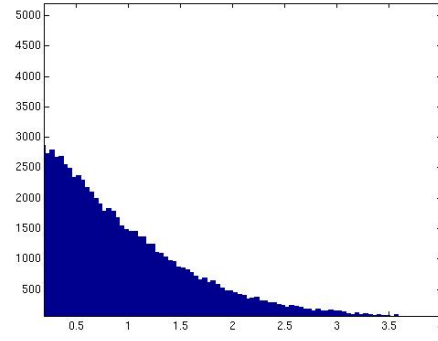


Figure 2. The histogram of mean delta pitch (MDP) calculated over 100,000 vowels.

of loudness is used, loudness can become a significant parameter to estimate prominence, and f0 might not play the big role it is usually though to. In Heldner (2001), it is found that spectral emphasis in Swedish correlates to prominence.

In this study we estimate loudness using the ITU (ITU- R1077, 2003) recommended loudness measure, which was recently evaluated to be a highly reliable perceptual measure by (Nygren (2009)), 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 vowel. Figure 3 presents the histogram of loudness using 100,000 vowels. As shown in the figure, the loudness distribution may follow a Gaussian shape, but since the literature does not suggest that the decreased vowel loudness increases or decreases the perception of prominence, loudness is modeled in our prominence model as a half-Gaussian distribution, (μ_l, σ_l^2) with μ_l is the vowel’ mean loudness, and σ_l^2 as the standard deviation, where loudness values below μ_l get 0 prominence. Equation (2) for loudness prominence level function becomes:

$$\Pr_x = 1 - \begin{cases} \frac{L(x|\mu_l, \sigma_l^2)}{L(\mu_l|\mu_l, \sigma_l^2)} & x \geq \mu_l \\ 1 & x < \mu_l \end{cases} \quad (6)$$

Where L is the half-gaussian likelihood function.

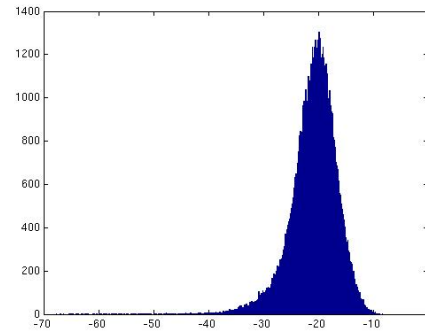


Figure 3. Histogram of vowel loudness using 100,000 vowels.

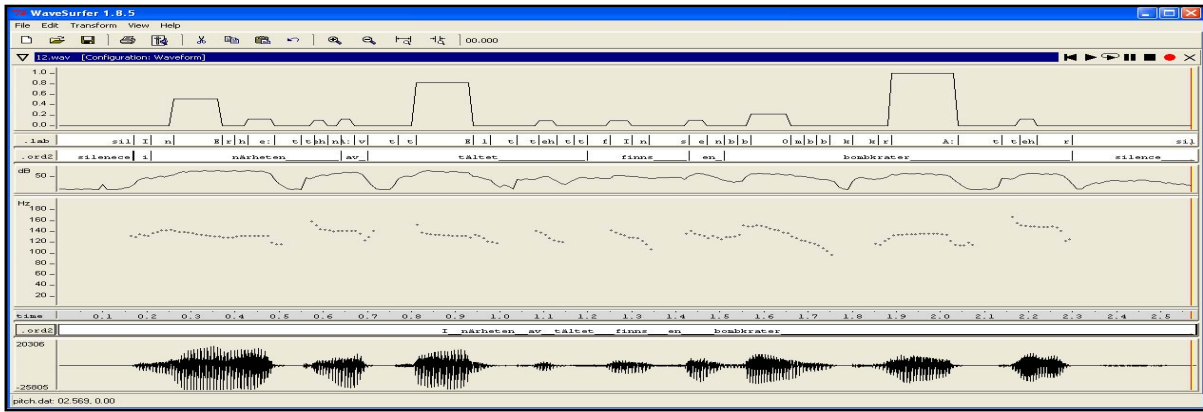


Figure 2: A plot visualizing the prominence level (top) with the acoustic parameters and a phonetic alignment of the sentence: “I närheten av tältet finns en bombkrater” using WaveSurfer.

5. Experiments

The way to deal with prominence segments in speech usually depends on the application, for example, detecting the most prominent syllable in a word, the most prominent word in a sentence, the most prominent syllable in an utterance, etc.

In this work, two experiments are conducted to investigate our proposed method. The first one takes advantage of a word level prominence annotation of read aloud speech produced by one speaker. The other evaluation uses a small test set annotated by a speech expert with focally-accented syllables.

Word Level Prominence

In (Strangert & Heldner 1995), 2 minutes of read aloud speech containing 250 words were annotated on a word level with prominence scale between 0 and 3 by 9 Native Swedish speech expert. The prominence level of each word was estimated using the mean annotations of all the experts for each word, and then scaled to the range [0, 1]. To estimate prominence using our method, we hypothesized that the highest prominence value of all the vowels (syllables) in the word represents the word prominence. In this case, by distributions of the duration, pitch, and loudness were estimated using the full set of data, and a normalized prominence value for each word is estimated. An MSE measure of this estimate resulted in 28.6% error. Figure 5 shows an example of the estimated and annotated normalized prominence level for an example sentence.

Focal-accent detection

50 utterances are selected from a speech corpus containing high quality one-speaker read speech designed for speech synthesis sound creation. One native speech expert marked focal-accent on a syllable level in all the utterances. This resulted in a variant number of focal-

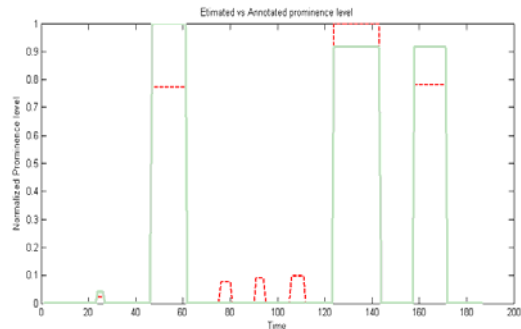


Figure 3. Automatic prominence estimation on an example sentence “affärsman och hans familj”. Dashed red line is the estimated value. Green line is original annotation.

accents marks per sentence with a sum of 120 markers in all the utterances. The prominence parameters distributions were adapted to these sentences, and a selection of a number of vowels with the highest prominence level is done, where this number matched the number of annotated focal accents in the file. This maps to a binary classification task with a prior knowledge of how many points lie in each class. On this small test set, a correct rate of 81.7% is achieved.

6. Conclusions

In Rosenberg (2009), 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 are 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 phonetic and segmental model defined over time, as, for example, in focally accented words. This implies many complications. On the first hand, labeling level of prominence on a segment basis shorter than a word, and limits the availability of training data. On the

other hand, it also limits the credibility of a prominence level over segments like vowels (or syllables) where the underlying prosodic information is not sufficient enough to capture a prominence perceptual level over a segment longer than that. This work presents a trial to build an unsupervised, predefined function of prominence over vowels, which shows through our experiments that statistically it is still plausible to estimate prominence as a continuous value (experiment 1) and prominence as a binary category (experiment 2).

Although the proposed method is evaluated using one-speaker read speech with enough data to estimate the prosodic parameters distributions, it is possible to extend the estimation of the model to in real-time, where the mean and standard deviation are calculated using the incoming signal with a decaying window, so it adapts to new speakers, differences in recording conditions (louder or softer signals), and to gender, which would be the upcoming extensions of this work.

Another important evaluation of this method, which has a potential to increase its accuracy, is to build a model of the prosodic parameters for each type of vowel as discussed before, which although might need more time for adaptation, but which would result in a more accurate estimate of the parameters compensating for the intrinsic differences in the (loudness, duration) of the vowels, and this would be plausible for tasks like annotating large corpora with prominence where there is enough data to build accurate distributions of the parameters.

References

- Bruce, G., & Granström, B. (1997). Prosodic modelling in Swedish speech synthesis. In Fant, G., Hirose, K., & Kiritani, S. (Eds.), *Analysis, Perception and Processing of Spoken Language. Festschrift for Hiroya Fujisaki*. (pp. 62-73). Amsterdam, The Netherlands: Elsevier Science B.V..
- Fant, G., & Kruckenberg, A. (1994). Notes on stress and word accent in Swedish. *STL-QPSR*, 35(2-3), 125-144.
- Fant, G., Kruckenberg, A., & Liljencrants, J. (2000). The source-filter frame of prominence. *Phonetica*, 57, 113-127.
- Heldner, M., & Strangert, E. (2001). Temporal effects of focus in Swedish. *Journal of Phonetics*, 29(3), 329-361.
- Heldner, M. (1998). Is an F0-rise a necessary or a sufficient cue to perceived focus in Swedish?. In *Nordic Prosody: Proc of the VIIIth Conference* (pp. 109-125).
- Heldner, M. (2001). Spectral emphasis as an additional source of information in accent detection. In Bacchiani, M., Hirschberg, J., Litman, D., & Ostendorf, M. (Eds.), *Prosody 2001: ISCA Tutorial and Research*

Workshop on Prosody in Speech Recognition and Understanding (pp. 57-60). Red Bank, NJ.

- Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3), 396-414.
- Nygren, P. (2009). *Achieving equal loudness between audio files - Evaluation and improvements of loudness algorithms*. Master's thesis, KTH CSC TMH.
- Rosenberg, A. (2009) Automatic Detection and Classification of Prosodic Events. PhD Thesis. Columbia University.
- Strangert, E and Helder, M (1995). Labelling of boundaries and prominence by phonetically experienced and non-experienced transcribers. *Phonum*. 33, 85-109.
- Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America*, 89, 1768.
- Wang, D., & Narayanan, S. (2007). An acoustic measure for word prominence in spontaneous speech. *IEEE Transactions on Audio Speech and Language Processing*, 15(2), 690.