Up to what level can acoustical and textual features predict prominence

Barbertje M. Streefkerk*, Louis C. W. Pols* and Louis F. M. ten Bosch**

*Institute of Phonetic Sciences (IFA)/ACLC, Univ. of Amsterdam
**Lernout & Hauspie Speech Products N. V., Belgium
barber@hum.uva.nl

Abstract

In this paper both acoustical as well as textual correlates of prominence are discussed. Prominence, as we use it, is defined at the word level and is based on listener judgments. A selection of useful acoustic input features is tested for classification of prominent words, with the help of Feed Forward Nets. We use spoken sentences from many different speakers, taken from the Dutch Polyphone corpus of telephone speech. For an independent test set of 1,000 sentences about 72% of the words are correctly classified as prominent or not. At the text input level we also developed an algorithm, using linguistic/syntactical features derived from text only, to predict prominence. The prediction agrees with the perceived prominence in 82.6% of the cases.

1. Introduction

In current speech technology applications the use of prosodic information is not without difficulties. In text-to-speech synthesis, accentuation (most of the time in terms of pitch accents) and phrasing can only rely on textual information [4]. Variation in strength of the boundaries and in prominence of the pitch accents is not taken into account in current speech synthesis systems. The simple distinction between accented and non-accented is not sufficient to make the intonation more natural. Attempts are made to introduce the notion of prominence [6] for speech synthesis purposes. The first step towards that is to predict the prominence based on available textual information. The second step is to transform the predicted prominence into acoustical features in order to incorporate the various prominence levels and to make them perceivable.

In most current speech recognition systems prosodic information is not used at all. However, a device that would indicate the degree of prominence of every word or of specific words would be helpful in several speech recognition applications. Prominence could then serve as an indication for islands of reliability, that can carry important and/or new information, such as a negation. Proper recognition of such words can be very advantageous. It could also be used in various ways in dialogue systems. Managing a dialogue frequently requires disambiguation of utterances, for instance of the type “in CAPable hands” versus “INcapable hands”.

The present project might help to incorporate prosodic features into speech recognition and into speech synthesis applications. For a set of 2,224 sentences, word prominence is marked by listeners and the acoustical as well as the textual correlates are investigated and tested for their predictability for prominence.

2. Speech material and initial prominence labeling

In this research project the speech material is taken from the Dutch Polyphone Corpus [3]. Phonetically rich newspaper sentences are spoken by people from all over the country. Each individual reads aloud five of these sentences which are then recorded over the telephone. From this material of over 5,000 speakers, we have randomly selected 1,244 sentences for training and an additional 1,000 sentences for testing, in which altogether 497 different speakers are involved. These sentences generally have a rather simple grammatical structure, with on average 10 words per sentence. However, the recording conditions vary a lot and huge speaker variability is created because people speak in their own home environment over the telephone.

2.1. Design of the training set

10 Naive listeners were asked to mark all prominent words of the 1,244 sentences of the training set. A subset of 50 sentences was presented twice, in order to get more information about the within- and between-listener agreement [7]. The individual marks of the listeners are binary (either 0 or 1), but the summed marks per word result in a prominence scale from 0 to 10. The agreement between and within the 10 listeners can be expressed by Cohen’s Kappa. On average κ = 0.5 with a standard deviation of 0.16.

Because this 10-points scale only suggests a high accuracy and furthermore depends on the actual number of listeners used, we simplified it, by means of a hierarchical cluster analysis, to a 4-points scale (0, I, II, III), which is further simplified to a to points scale putting 0 and I together as well as II and III. This merging of categories makes it more comparable with the test set. Zero indicates no prominence at all and III is the category of highest word prominence. See Table 1 for the distribution of these two combined classes in the training data.

Table 1: Absolute and relative number of words in the training set of 1,244 sentences belonging to the four prominence classes.

<table>
<thead>
<tr>
<th>Prominence</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 and I</td>
<td>7,818</td>
<td>59.6 %</td>
</tr>
<tr>
<td>II and III</td>
<td>5,301</td>
<td>40.4 %</td>
</tr>
<tr>
<td>Total</td>
<td>13,119</td>
<td>100 %</td>
</tr>
</tbody>
</table>
2.2. Design of the test set

For efficiency reasons, the test set of 1,000 sentences was processed in a slightly different way. From the initial group of 10 listeners only the one with the highest between (mean \( \kappa = 0.59 \)) and within (\( \kappa = 0.8 \)) agreement was chosen to mark the prominence of the words in this test set. So, for these words only a binary score of 0 (indicating no prominence) or 1 (indicating prominence) was available. Of the total of 10,330 words, 3,998 words were marked by this listener as prominent, this is 39%, which is comparable to the 40.4% for classes II and III in the training data, see Table 1.

3. Prominence classification based on acoustical features

We automatically segmented the speech material at the phoneme level by using a forced alignment with an HHM speech recognizer [8]. Since we suppose the text to be known, word boundaries, syllable boundaries and segment boundaries are then also available and measurements can be done at all these levels. Because the prominence marks are at the word level and acoustical features are not always extracted at the word level but for vowels and syllables as well, we limit ourselves in the case of polysyllabic words the lexically stressed syllables only. In general the following features are supposed to have an influence on word prominence: \( F_0 \), intensity, duration, and perhaps spectral quality, see also [1].

The 7 basic features that we use are: vowel and syllable duration, vowel intensity, \( F_0 \) range per word and per syllable, and \( F_0 \) median value per word and per syllable. Beyond that, 7 additional features are used, namely the sentence speaking rate, the overall intensity of a sentence, and the median \( F_0 \) of a sentence, plus some normalized features namely, vowel duration normalized for the intrinsic duration and vowel intensity normalized for the intrinsic intensity.

3.1. Acoustical features

It is to be expected that the stressed syllables in prominent words, and thus also the vowels, are louder, longer and show more pitch variation than in the non-prominent words. Furthermore, it is certainly worthwhile to explore how much discriminative power can be gained by normalizing, among other things, for intrinsic vowel duration and for speaking rate.

As was already shown by Pols et al. [5] for the TIMIT sentences, also these Polyphone sentences show substantial overall variation in sentence speaking rate. This speaking rate \( (r) \) is defined as the average normalized phoneme duration \( (\tau) \) per sentence.

\[
\tau = \frac{1}{N} \sum_{i=1}^{N} \tau_i \quad \tau = \frac{d - \mu}{\sigma}
\]

Zero implies an average speaking rate, a positive value implies a slow rate and a negative value a fast rate. The actual variation in speaking rate over all sentences can be seen in Fig. 1. The mean vowel duration as a function of the sentence speaking rate. The two curves represent the most (scale III) and the least (scale 0) prominent vowels.

3.2. Classification with feed forward nets

By using various combinations of the basic acoustic features (normalized between 0 and 1) at the input level, and by using a hidden layer with a variable number of nodes, we trained a feed forward net that had a binary output (prominence 0 or 1). The network was trained with equal (randomly selected) numbers of scale 0 plus I and scale II plus III data, in order to make this training set more comparable with the test set. In total a set of maximally 14 features is used to train these feed forward nets for binary prominence classification.

3.3. Results

So far the first preliminary results show a percentage correct prominence classification of about 79% for the training data and about 72% for the independent test data. This holds for using all 14 features at the input. Single basic subsets of features, such as related to duration, intensity or pitch only, show somewhat lower performance, but the asymmetry in the test set performance for the prominence 0 and prominence 1 classes then becomes rather high. These and other phenomena require further analyses.
4. Prominence prediction based on textual information

In synthesis applications it would improve the speech quality and the communicative function of the speech material generated, if we were able to properly predict the word prominence from textual information only. In most present-day text-to-speech systems one does not get much further than giving all content words a pitch accent. In the present paper we explore the possibilities of using several textual correlates to properly predict prominence. Since it is rather ambitious to extract meaning from a given sentence in a given context, we limit ourselves to correlates that can be derived automatically from text, such as: POS, number of syllables, position of words in the sentence and co-occurring word classes, such as the Adjective-Noun combination. Word classes were assigned automatically for the test and training set by a memory-based Part-of-Speech (POS) tagger [2], which compares a particular word in a particular context with a most similar case in memory. Comparing its performance with hand derived POS labels for the training set of 1244 sentences shows 92% correspondence in labels.

4.1. Textual features

11 Word categories are distinguished: Articles, Conjunctions, Prepositions, Pronouns, Auxiliary Verbs, Verbs, Numerals, Adverbs, Adjectives, Nouns, and Negations. Table 2 shows the frequency of occurrence of these word categories in the training data, as well as the mean prominence score (between 0 and 10) and its standard deviation per word class. The word classes are ordered from least to highest average prominence. It is clear that, next to Nouns, also Numerals, Negations, and Adjectives receive high prominence.

From Figure 3 it can be seen that generally the prominence of the word classes Noun, Adjective, Numeral and Negation is much higher if these words occur in the first position of the sentence rather than elsewhere in the sentence. This is another element that is taken care of in developing the rules.

![Figure 3: Distribution of the prominence of the word classes Noun, Adjective, Numeral, and Negation either at the initial position of the sentence, or elsewhere.](image)

![Figure 4: This is a comparison of the distribution of the prominence over the Nouns, when they are either preceded by an Adjective or in all other positions.](image)

The same is true for the phenomenon illustrated in Figure 4. This shows that Nouns preceded by an Adjective generally have a substantially lower prominence than all other Nouns. These and other characteristics in our sentence material have been used to optimize our algorithm for predicting prominence.
4.2. Algorithm to predict prominence

From the regularities found in our data, we have so far derived the following rule set:

- **rule I**: every content word receives one mark
- **rule II**: every word from the classes \{Noun, Adjective, Numeral, Negation\} receives an additional mark
- **rule III**: every polysyllabic word from the classes \{Pronoun, Verb, Adverb\} receives an additional mark, and every word from the classes \{Noun, Adjective, Numeral, Negation\} receives once more an additional mark
- **rule IV**: the first content word in the sentence receives an additional mark
- **rule V**: every Noun preceded by an Adjective is decreased by one mark

4.3. Results

Table 4 shows the results of applying these rules to the training sentences. The second column gives the frequency of occurrence of each prominence mark. The next columns indicate the related mean values and standard deviations of the actually perceived prominence. Generally there is a good overall relation between perceived and predicted prominence.

<table>
<thead>
<tr>
<th>Predicted Prominence</th>
<th>Total occurrence</th>
<th>Prominence Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5,796</td>
<td>0.40</td>
<td>1.36</td>
</tr>
<tr>
<td>1</td>
<td>1,480</td>
<td>3.15</td>
<td>3.40</td>
</tr>
<tr>
<td>2</td>
<td>2,430</td>
<td>3.84</td>
<td>3.09</td>
</tr>
<tr>
<td>3</td>
<td>2,673</td>
<td>5.85</td>
<td>2.77</td>
</tr>
<tr>
<td>4</td>
<td>740</td>
<td>7.50</td>
<td>2.25</td>
</tr>
<tr>
<td>Total</td>
<td>13,119</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The actual test of course should be done with the independent test set, which however has the drawback that the perceived prominence level is only defined in a binary way. For the results, see Table 5. This table is multi-interpretable, since one can normalize both horizontally as well as vertically. Only 13.8% (553/3,998) of the words labeled by the listener as prominent are actually marked with the highest prominence level 4, which seems to be a very low score. However, 92.6% (553/597) of the words predicted with prominence mark 4 are also perceived as prominent, which certainly is a very high score. Simplifying this table to a (2 x 2) table by combining the marks 0 and 1 as well as the marks 2 to 4, leads to 76.5% and 88.7% correct prediction for perceived prominence 0 and 1, respectively. More research is required to optimize the predicted prominence marks and to transform them into adequate acoustical features and into an appropriate synthetic speech quality.

<table>
<thead>
<tr>
<th>Perceived Prominence</th>
<th>Predicted prominence mark</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4,001 841 930 516 44</td>
<td>6,332</td>
</tr>
<tr>
<td>1</td>
<td>180 272 1,284 1,709 553</td>
<td>3,998</td>
</tr>
<tr>
<td>Total</td>
<td>4,181 1,113 2,214 2,225 597</td>
<td>10,330</td>
</tr>
</tbody>
</table>

5. Concluding remarks

This study has shown that it is possible for difficult text input material (individual sentences from the Dutch Polyphone corpus), to predict with reasonable accuracy the word prominence from linguistic/syntactical features based on the isolated sentence text input only. We were able to test the accuracy of this prediction since these sentences were also available in spoken form and a listener had marked all prominent words in that spoken material.

This also allowed us to run the opposite test, namely to predict the word prominence, based on acoustical features extracted from the speech signal only. In section 3 we have shown that a simple feed forward net with one hidden layer, can predict this word prominence rather well, if fed with the appropriate acoustical features. We are in the process of testing how much it helps to normalize certain features, for instance for intrinsic vowel duration and vowel intensity, as well as for average sentence pitch and for sentence speaking rate.

6. References