5

NEURAL NET CLASSIFICATION OF PROMINENCE WITH ACOUSTIC INPUT FEATURES[1]

Abstract

The main topic of this chapter is the classification of word prominence, exclusively based on acoustic input. Prominence is classified by means of feed-forward networks. This chapter includes a brief description of such neural networks. The input is chosen from the acoustic correlates as described in chapter 4. In that chapter two applications were suggested, namely a sentence disambiguator and a word prominence indicator. The consequences of such applications in terms of training factors are examined in this chapter. We obtained a prominent / non-prominent classification accuracy of 82% for the development test set and 79% for the independent test set.

[1] Parts of this chapter were published in Streefkerk et al. (1998), Streefkerk et al. (1999 a), Streefkerk et al. (1999 b), and Streefkerk et al. (2001).

5.1 Introduction

In this chapter we deal with the classification of prominence based exclusively on acoustic input features. The subsections of this introduction give a general description of a neural network and two examples of possible applications for prominence classification. These applications bring us closer to the question: what do we want the network to do. Simple feed-forward networks are used to recognize prominence with selective data. These exemplary data are presented to the neural net during the training phase and indicate which features are especially important for achieving correct classification. The input pattern consists of characteristic acoustic features, as described in chapter 4. Based on such characteristic input patterns, a trained neural net predicts whether or not words are prominent. The advantage of neural networks is that no specific knowledge has to be expressed in rules; instead, the knowledge is based on the training data sample. Other classification techniques are possible, for instance, a linear discriminant analysis (LDA). In such an analysis, however, only linear relationships can be found. Representing complex relationships may be needed for our classification problem. Examples for complex data classification with neural networks are described in Weenink (1991) and in Lippmann (1987). For our analyses we try to keep the topology of the net as simple as possible.

5.1.1 How feed-forward networks work

Feed-forward networks consist of units (nodes) and activation functions. The basic units of neural networks are the nodes; several nodes are grouped into layers (see figure 5.1). A learning algorithm allows the neural network to learn a certain task by adjusting the weights. The layer(s) between the input and the output layer are called hidden layer(s), (see figure 5.1). The formula below expresses the relationship of the output with the output of nodes in a previous layer (x_i), weights associated with the connections (w_i) and a threshold (θ) of the node. With these variables the output (y) is calculated, via an activation function (f), for instance a sigmoid function. Each node converts the pattern of incoming activities into one single activity. This single activity ('output') is passed on to the other connected nodes in the next higher layer. The activation function typically falls into one of three categories:

- a linear function (the output is proportional to the total weighed input),
- a threshold function (the output is set at one of two levels, depending on whether the total input is greater than or less than some threshold value), or
- sigmoid functions (the outgoing activity varies non-linearly with the input (weighted)).

Generally the activation function can be expressed as:

$$y = f\left(\sum_{i=1}^{N} w_i x_i - \theta\right)$$

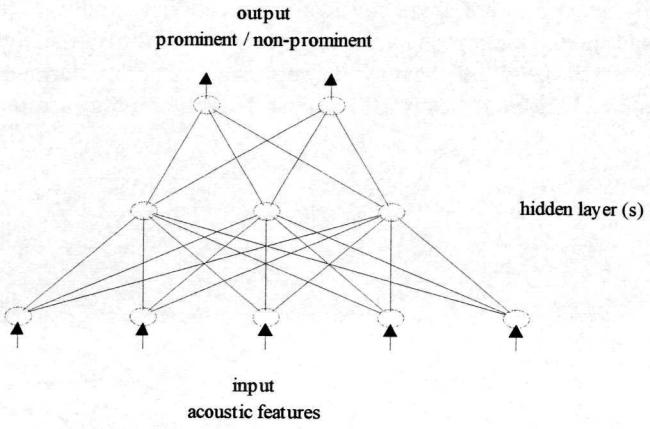


Figure 5.1: A possible topology of a neural network.

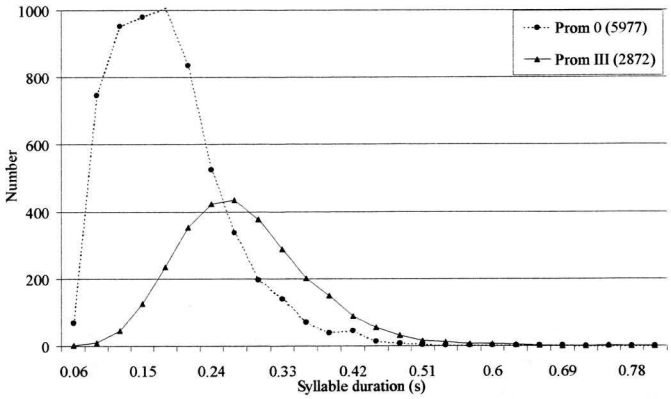
The behavior of a neural net depends on the weights (w_i) of the connections between the nodes, on the threshold (θ) of the node, and on the activation function (f). The weights and the threshold values can be adjusted during training; the activation function is usually fixed during training.

5.1.1.1 General training procedure

The main steps of the supervised learning process of the net can be described as follows: The acoustic features are presented at the input layer. The values of the features are sent via the weighed connections of this input layer and via weighed connections of one or more hidden layers to the output layer. The computed output is compared with the desired output values. Based on this comparison, the difference is calculated and the weights and the threshold values are subsequently adjusted (e.g. by the back-propagation algorithm, Rumelhart et al., 1986).
In the back-propagation algorithm the weights are adjusted per training pair (input features plus desired output) by using a feedback step. This makes the adjustment rather sensitive to the sequence in which the training pairs are presented to the net. A more sophisticated algorithm is the conjugative gradient method (Press et al., 1992). This training algorithm calculates the difference between the desired output and the calculated output for the whole training set, and then starts to adjust the weights and the threshold of each node. This whole procedure is called an iteration step. Such a training procedure makes training less sensitive to local minima. In this study we use the more sophisticated conjugative gradient method.

If enough examples are presented to the net, the net is able to generalize over the various characteristic input features. For new (unseen) data this trained net is able to predict which label belongs to the presented input pattern. The net bases its knowledge on the examples it has seen before in the training session.

a)



b)

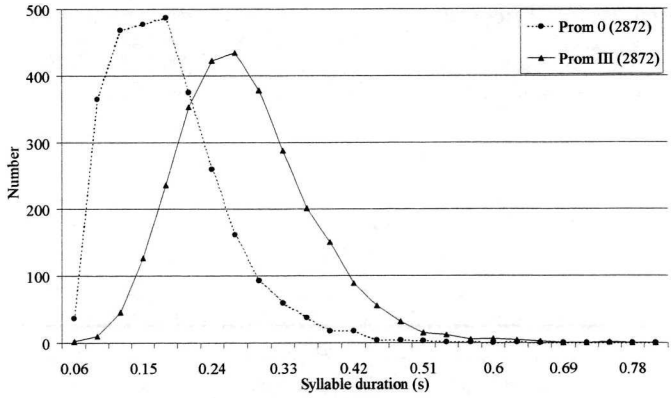


Figure 5.2: Two different distributions of the syllable duration for prominence class 0 and III. Graph a) displays the distribution based on the actual number of occurrences as found in the set of 1244 sentences. Graph b) displays two distributions of which the sizes have been made equal by random selection.

5.1.2 Distributions of prominence classes and the relationship to applications

Two applications of prominence classification have already been mentioned in chapter 4. One application is a prominence-indicator that measures the amount of prominence that each word carries. This classification is generally based on a biased distribution of prominent and non-prominent words (unequal numbers of prominent and non-prominent words). See figure 5.2a for an example of such a biased distribution concerning syllable duration. A classifier for the prominence of all words occurring in the sentences has to take into account this bias in the training data.
The other application is the disambiguation of two possible interpretations of a sentence, for instance in *uitsluitend VOOR instappen (only get on at front)* versus *uitsluitend voor INstappen (only for getting on)*. In such a disambiguation task there are two words (syllables) involved and it must be decided which of the two words is the most prominent one. To answer this question, the classification has to be based on an unbiased distribution (= equal number of prominent and non-prominent words). Figure 5.2b is an example of this.
To cover both options in the analyses, our training and test data will be split up into a biased and an unbiased set. We are aware of the fact that neural networks are able to account for the prior probability of prominence in the data, but by using the unbiased and biased distributions for training and testing the neural networks are optimally trained for such a task, as mentioned above. So, it makes sense to use biased and unbiased sets as different training conditions.

5.2 Prominence recognition with neural networks

Before going into detail about the contribution of each individual acoustic feature and the performance of an 'optimal' neural network, a brief description of the acoustic features and their pre-processing is given in the following subsections.

5.2.1 Acoustic input features

In chapter 4 the main acoustic features were described in detail and a number of results of this analysis of useful acoustic features for prominence classification were presented. These features are complemented with overall features such as the median F_0 of the sentence. The total set of acoustic features is a set of twelve features as displayed in figure 5.3 and given below.

1. vowel duration;
2. vowel duration normalized for intrinsic vowel duration;
3. sentence speaking rate;
4. vowel intensity normalized for the overall intensity of the given sentence;
5. vowel intensity (sentence normalized) normalized for the intrinsic vowel intensity;
6. overall intensity per sentence;
7. syllable duration;

8. median F_0 per syllable;
9. range of F_0 per syllable;
10. median F_0 corrected for the median F_0 per sentence;
11. median F_0 of the sentence;
12. range F_0 per word.

In figure 5.3 this set of twelve features is displayed as input for a neural network.

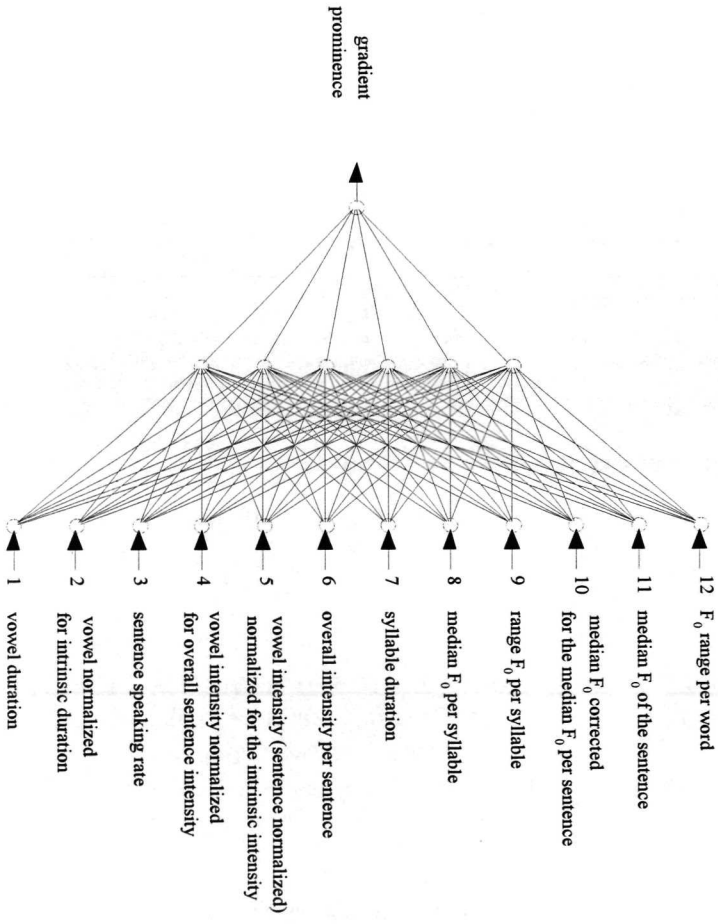


Figure 5.3: An example topology of a neural network, with input features and a single output node, for gradient prominence prediction.

5.2.2 Pre-processing of the input features

The network algorithm implemented in PRAAT prefers inputs between 0 and 1. Therefore all input features are scaled between 0 and 1. This adaptation is a specific requirement for this software and does not apply for neural nets in general. This scaling is performed in the following way: Firstly, per acoustic feature the 99% percentile (max) and the 1% percentile (min) are determined for all training and test data together. (See for more details Appendix A 5.1.) All values occurring beyond these borders are set to 1 and 0, respectively. (Wrong measurements only mildly influence the scaling factor.) The values are scaled by the following formula:

$$x_{scaled} = \begin{cases} 0 & x < \min \\ 1 & x > \max \\ \dfrac{x - \min}{\max - \min} & \text{else} \end{cases}$$

5.2.2.1 Correlation

The set of twelve features obviously contains features with correlated information. For instance, the median F_0 features in corrected (10) and uncorrected (8) form may contain such dependent information. The correlation matrix of all features used is given in table 5.1.

Table 5.1. The correlation matrix of the twelve acoustic features. The acoustic features are indicated with a feature number; the feature description is given in figure 5.3.

Feature number	1	2	3	4	5	6	7	8	9	10	11	12
1	-	0.82	0.23	0.19	0.01	0.12	0.57	0.08	0.26	0.00	0.10	0.29
2	-	-	0.29	0.06	0.05	0.14	0.40	0.11	0.22	0.01	0.14	0.21
3	-	-	-	0.03	0.03	0.24	0.27	0.10	0.12	-0.01	0.15	0.12
4	-	-	-	-	0.90	-0.02	0.28	0.12	0.14	0.18	-0.01	0.17
5	-	-	-	-	-	-0.04	0.09	0.13	0.08	0.20	-0.01	0.09
6	-	-	-	-	-	-	0.08	0.24	0.09	-0.01	0.32	0.09
7	-	-	-	-	-	-	-	0.04	0.37	0.01	0.05	0.34
8	-	-	-	-	-	-	-	-	0.07	0.65	0.71	0.07
9	-	-	-	-	-	-	-	-	-	0.10	0.01	0.78
10	-	-	-	-	-	-	-	-	-	-	-0.06	0.09
11	-	-	-	-	-	-	-	-	-	-	-	0.01
12	-	-	-	-	-	-	-	-	-	-	-	-

The correlations calculated on the original data group into four parts; first high correlations between features that are based on each other such as the vowel duration (1) and the normalized vowel duration (2). Their correlation is 0.82. The second group concerns lower correlations between features that contain similar information such as vowel duration (1) and syllable duration (7). Their correlation is 0.57. The third group shows positive but rather low correlations such as vowel duration (1) and F_0 range per word (12), which correlates 0.29. Fourth are those features that hardly correlate at all, e.g. sentence speaking rate (3) and vowel intensity sentence normalized (4). Their correlation is 0.03.
Some features are included in our analysis simply because we want to test what the effect of various corrections is. It will be tested further on in this chapter whether two highly correlating features should be left out as input features in order to obtain a better performance. As a first step we will train with all twelve input features, neglecting the fact that certain features correlate highly, because in brute force research of, for instance Kießling (1996), the use of as many features as possible leads to high correct classification rates.

5.2.3 Design of the training and testing data

There are several factors that influence the performance of the net. The first is feature representation. Second, the number of hidden nodes in the hidden layer also influences the performance of the net. This number is directly related to the degrees of freedom the neural network has to adjust itself to the training material. Third, the number of iterations will also influence the results. Finally, the distribution over the different prominence categories, as mentioned in section 5.1.2, influences the results. The output needs further specification. One could train and test with discrete output; this is done with separate nodes for each prominence class. Alternatively, one could train and test with continuous output; this requires only one output node. Such a single node with linear output (or with sigmoid) is enough to predict a gradient prominence scale.

The general structure of the test and training set was described in chapter 2, so a summary here will suffice. As described there, the training set consists of 1244 sentences marked for prominence by ten listeners, which resulted in cumulative prominence marks between 0 and 10. By means of a hierarchical cluster analysis the scale was reduced to four classes, namely 0, I, II, III. For simplicity's sake the discrete output was set to two possibilities, namely 'non-prominent' (containing prominence class 0 and I) and 'prominent (containing prominence class II and III). Figure 5.4 gives more details about the distribution of training and test data.
Only one selected 'optimal' listener judged the 1000 sentences of the Independent Test set. With this test set it can be independently tested whether the neural network behaves similarly to one of the listeners. This can only be tested for the binary prominent / non-prominent distinction. We decided to train separate neural networks for prominence degrees i.e. gradient prominence prediction as well as for binary prominence prediction.

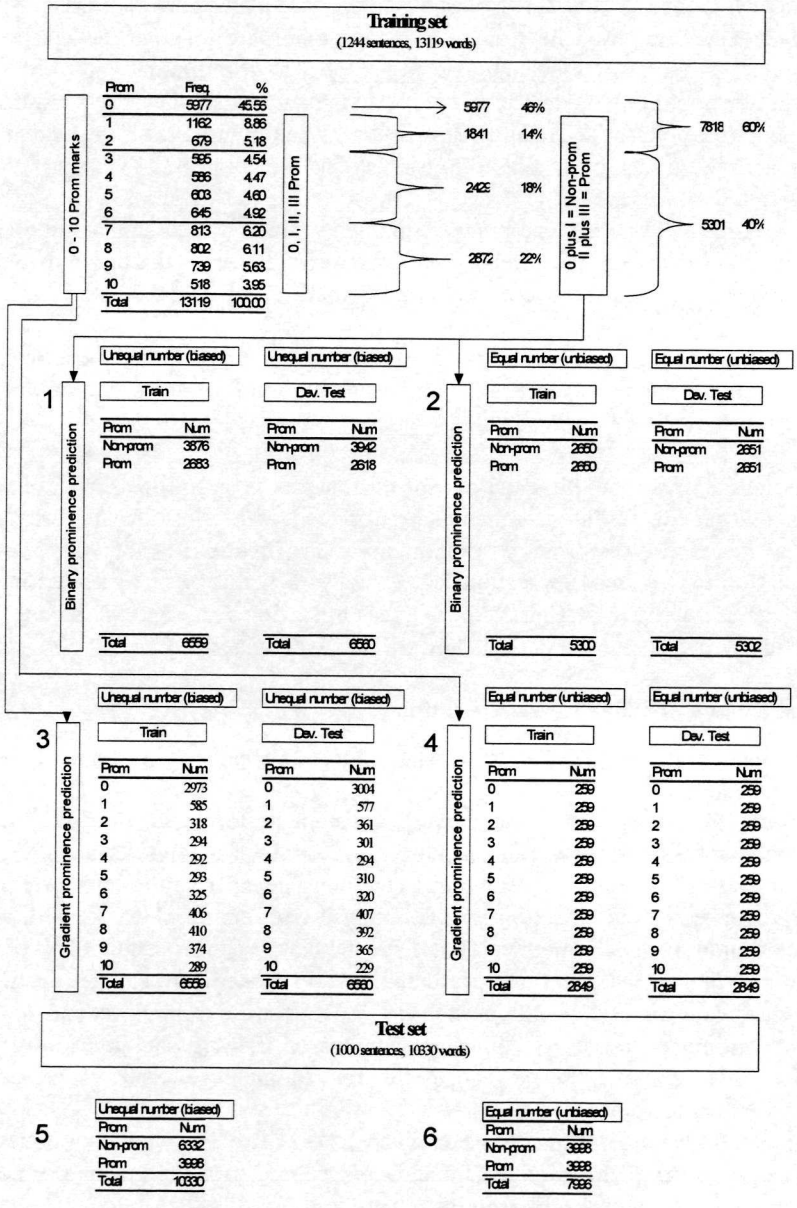


Figure 5.4: Diagram of the Test and Training sets, their distribution over biased and unbiased Training sets and Development test sets, and their distribution over binary and gradient prominence classes. In the Independent Test set only binary categories are used.

In order to train and test the 0-10 gradient prominence scale, the original training set of 1244 sentences with 13119 words, was randomly divided into two equal parts; once for biased and once for unbiased training. See also figure 5.4 dataset 3) and 4). The first part was used for training purposes (Training set) and the second part to test during training (Development test set). The performance of the gradient prediction could only be tested by the Development test set, because a gradient scale was not available for the Independent Test set. (See figure 5.4 for the exact numbers of data in these sets for the biased 5) training and the unbiased 6) version.)

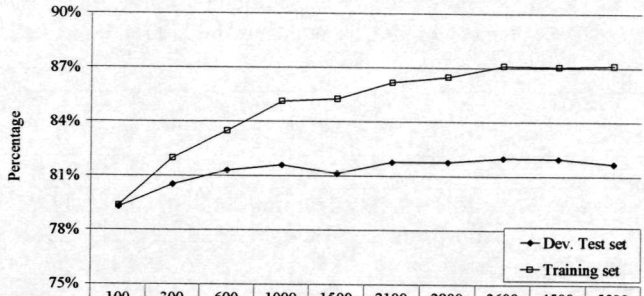
In order to train and test the binary prominence prediction, a similar subdivision of the training data was performed. Our data were randomly divided into two equal parts; one for biased and one for unbiased training. See figure 5.4 dataset 1) and 2).

As for gradient prominence prediction there is also a Training and Development test set. An Independent Test set, one for biased and one for unbiased condition, figure 5.4 dataset 5) and 6), was used to test binary prominence prediction.

The outline of the experimental part of this chapter is as follows. First we will deal with the general results of a neural network fed with all twelve features. Section 5.2.4 describes the binary prominence classification. Gradient prominence prediction is discussed in section 5.2.5. Analyses with a set of selected individual features are given in section 5.2.6, whereas in section 5.2.7 and 5.2.8 combinations of features are the topic. A conclusion will be given in section 5.3.

5.2.4 Binary prominence classification

This section is concerned with dataset 1) and 2) (figure 5.4): binary prominence classification.
In order to get an idea about the classification performance first of all a linear discriminant analysis was run on these data of which the results are given in the upper part of table 5.2. With an LDA only linear relationships are used for classification, so classification results obtained with neural networks with a hidden layer should always be higher. Therefore, the classification results of the LDA are used as a bottom indication. As presented in table 5.2 the correct classification using unbiased data (dataset 2) in figure 5.4) is 77.01% for the training set and 76.05% for the development test set. Training with the biased training data gives lower correct prominence classification, 76.86% for the training set and 75.71% for the development test set.
The neural nets are designed in such a way as to create are two output nodes; one is active when the features of the input vector belong to a prominent word, and the other one is active for a non-prominent word.
Several networks under biased and unbiased conditions were trained with the number of hidden nodes varying from 2 to 18, whereas the number of iterations differed also. Degrees of freedom vary from 32 with a 12-2-2 net up to 272 with a 12-18-2 net. The net decides by the so-called 'winner-takes-all' criterion. Figure 5.5



	100	300	600	1000	1500	2100	2800	3600	4500	5500
Dev. Test set	79.28%	80.52%	81.29%	81.59%	81.14%	81.77%	81.77%	82.01%	81.97%	81.69%
Training set	79.37%	81.98%	83.49%	85.13%	85.27%	86.18%	86.50%	87.11%	87.05%	87.13%

Figure 5.5: Correct recognition percentages of the Training and Development test sets classified with a neural network with the topology of 12-14-2 for the distinction of prominence non-prominence (unbiased training).

presents classification results on the Training and the Development test set with an increasing number of training iterations. This example is taken from table 5.2 and deals with the net with 14 hidden nodes. Whereas the curve for the training data steadily increases, the curve for the development test data starts to decrease somewhere around 3600 training iterations. The number of training iterations needed also depends on the training algorithm. At this turning point the net is sufficiently trained but at the same time is not too closely adjusted to the training data. The results, presented in table 5.2, are selected according to this criterion. Only results of fully trained nets are presented. The recognition rates are presented as overall results and results that have been separated for prominent and non-prominent recognition. Table 5.2 gives the results of several neural nets for the Training set (1, 2 in figure 5.4) and for the Development test set (1, 2 in figure 5.4). The percentages correct prominence classification with neural networks are always higher than classification with an LDA. The correct classification on the training data also are always higher than on the development test set data. Neural networks with more than 6 hidden nodes in their hidden layer give better performances than with less hidden nodes (see table 5.2).

Unbiased training (equal numbers) achieves better recognition results for prominence classification (around 82%) than for non-prominence classification (around 77%). The features describing the prominent words may be more clearly defined than the features describing non-prominent words. The opposite is true for the biased trained neural networks. There the non-prominence classification (around 82%) achieves higher recognition results than the prominence classification (around 76%). This is explained by the fact that a neural net trained with unequal numbers is biased to recognize the more frequently occurring non-prominent words. This

Table 5.2: Results (percent correct classification) of several neural networks with a variable number of hidden nodes in the hidden layer, trained with biased und unbiased data.

		Unbiased Training			Biased Training		
		Non-prom	Prom	%	Non-prom	Prom	%
Linear discriminant analysis (LDA)	All	Training set		77.01	Training set		76.86
	Non-prom	1996	654	75.32	2986	956	75.75
	Prom	564	2086	78.71	562	2056	78.53
	All	Dev. test set		76.05	Dev. test set		75.71
	Non-prom	1960	691	73.93	0.74355	994	74.87
	Prom	579	2072	78.16	0.77674	2084	78.79
Num hidden nodes							
2	All	Training set		**79.79**	Training set		**78.82**
	Non-prom	1993	657	75.21	3135	741	80.88
	Prom	414	2236	84.38	648	2035	75.85
	All	Dev. test set		**78.06**	Dev. test set		**78.09**
	Non-prom	1977	674	74.58	3161	781	80.19
	Prom	489	2162	81.55	656	1962	74.94
4	All	Training set		**81.13**	Training set		**80.99**
	Non-prom	2056	594	77.58	3208	668	82.77
	Prom	406	2244	84.68	579	2104	78.42
	All	Dev. test set		**79.80**	Dev. test set		**79.77**
	Non-prom	2042	609	77.03	3238	704	82.14
	Prom	462	2189	82.57	623	1995	76.20
8	All	Training set		**82.49**	Training set		**81.06**
	Non-prom	2083	567	78.60	3218	658	83.02
	Prom	361	2289	86.38	584	2099	78.23
	All	Dev. test set		**80.33**	Dev. test set		**79.85**
	Non-prom	2050	601	77.33	3234	708	82.04
	Prom	442	2209	83.33	614	2004	76.55
10	All	Training set		**84.26**	Training set		**83.34**
	Non-prom	2157	493	81.70	3279	597	84.60
	Prom	341	2309	87.49	496	2187	81.51
	All	Dev. test set		**81.25**	Dev. test set		[illegible]
	Non-prom	2059	592	77.79	3228	714	81.89
	Prom	402	2249	85.08	591	2027	77.43
14	All	Training set		**86.66**	Training set		**80.81**
	Non-prom	2232	418	84.67	3227	649	83.26
	Prom	289	2361	89.54	610	2073	77.26
	All	Dev. test set		[illegible]	Dev. test set		**79.91**
	Non-prom	2114	537	79.73	3254	688	82.55
	Prom	417	2234	84.28	630	1988	75.94
18	All	Training set		**82.55**	Training set		**81.26**
	Non-prom	2106	544	79.47	3247	629	83.77
	Prom	381	2269	85.62	600	2083	77.64
	All	Dev. test set		**80.20**	Dev. test set		**79.54**
	Non-prom	2068	583	78.01	3256	686	82.60
	Prom	467	2184	82.38	656	1962	74.94

shows that the distributions of prominent and non-prominent words play a role in the classification process.

In the unbiased case the best overall performance of prominence classification is 82.01% on the Development test set. This result was achieved in a neural network with 14 nodes in the hidden layer. The best performance for a biased trained network (10 hidden nodes) resulted in an overall recognition rate of 80.11%.

5.2.4.1 Testing with the Independent test set

The best nets based on the general results of table 5.2 were selected for independent testing. For the biased condition this is net 12-10-2, and for the unbiased condition it is net 12-14-2. These 'optimal' neural networks were used to mark the words in the 1000 sentences of the Independent Test set (figure 5.4 dataset 5 and 6). These prominence-marking results can easily be compared to the marks of the one 'optimal' listener who marked these words for prominence. The results are presented in table 5.3 and 5.4 in the form of a crosstable. Table 5.3 concerns the condition of the unequal numbers of prominent and non-prominent words (biased data) in the 1000 sentences of the Test set. The input data in table 5.4 deal with the unbiased case. The number of non-prominent words is then randomly reduced to the same number of available prominent words, which is 3998 (see condition 6 in figure 5.4).

Table 5.3: The correct recognition rates of prominence classification of the Independent Test set (biased data, condition 5 in figure 5.4). The total numbers as well as the percentages are given for the networks trained under unbiased and biased condition; the networks with the topology of 12-14-2 and 12-10-2 were optimal.

	Unbiased training			Biased training		
	Non-prom	Prom	%	Non-prom	Prom	%
	Test set (biased)		77.1	Test set (biased)		78.9
Non-prom	4907	1425	77.5	5232	1100	82.6
Prom	942	3056	76.4	1079	2919	73.0
Measure of agreement (κ)			0.53			0.57

First, we give a description of table 5.3. Percentages of prominence and non-prominence recognition rates are 77.5% and 76.4%, respectively, for the unbiased training condition (on average 77.1% correct). Non-prominent recognition is 82.6% for the biased training condition. This is comparable to the results of the development test set. Training with biased data gives better results on non-prominent recognition; the prominent recognition drops to 73.0%. A biased trained net performs best (78.9%) on the Independent Test set. The better performance of the biased trained net is as expected, as the distribution of the prominence marks in the 1000 sentences of the Independent Test set coincides with the distribution in the

training material. The Test set contains 10330 words; 3988 of these are marked as prominent, which is 39%. The remaining 61% is non-prominent. The 1244 sentences used for the training and development test set contain 13119 words, of which 7818 are treated as prominent, which is 40% of the total number of words.

When the bias in the data of the Test set is removed, i.e., when equal numbers of prominent and non-prominent words are presented, the performance of the net 12-10-2 decreases to 77.78% correct classification (see table 5.4). The performance of the unbiased trained net 12-14-2, tested with unbiased data of the Test set (condition 6 in figure 5.4) is 77.05% correct. Training and testing under the same biased or unbiased conditions do not give consistently better classification results. In table 5.3 the results are better when trained and tested with biased data, but contrary to this in table 5.4 the results for training and testing with unbiased data shows less percentages correct than trained with biased data and tested with unbiased data.
The between-listener agreement (section 2.4.1.2) expressed in Cohen's Kappa κ, was calculated for the results of the neural network and the listener who marked all 1000 test sentences. Kappa values are 0.53 (unbiased training) and 0.57, (biased training), see table 5.3. In an unbiased Test set (table 5.4) these Kappa values hardly differ. Similar values (on average $\kappa = 0.50$; Std. Dev. = 0.16) were measured for the between-listener agreements see section 2.4.1.2. This means that the neural network behaves similarly to any listener, and that the differences in prominence classification are as accurate as the prominence classification of any naive listener. The performance of the net is indistinguishable from any listener.

Table 5.4: This table presents the recognition rates of prominence classification on the Independent Test set (unbiased data, in figure 5.4 condition 6). The total numbers as well as the percentages are given for the 'optimal neural networks' trained under unbiased and biased condition with the topology of 12-14-2 and 12-10-2, respectively.

	Unbiased training			Biased training		
	Non-prom	Prom	%	Non-prom	Prom	%
	Test set (unbiased)		77.05	Test set (unbiased)		77.78
Non-prom	3105	893	77.66	3300	698	82.54
Prom	942	3056	76.44	1079	2919	73.01
Measure of agreement (κ)			0.54			0.56

5.2.4.2 Summary and conclusion

For the binary prominence classification the following results have been reached: 82% correct classification on a Development test and 79% correct on an Independent Test set. The performance may be accurate enough to allow sentence disambiguation to be done by such a classifier, especially if one keeps in mind that

any listener is indistinguishable from our 'optimal' neural network (79% correct, κ = 0.57, table 5.3) fed with acoustical information only. Considering that the listener has both acoustical and linguistic information, a combination of both acoustic and linguistic input features may further improve the classification.
Comparing correct recognition rates of an LDA with neural networks shows that apparently non-linear relationships exists between acoustic input features and prominence.
With an LDA only linear relationships can be used for prominence prediction, however, with neural networks with a hidden layer higher order relationships can be exploited. The prominence classification with a neural network appears to be always better than the classification with an LDA.

5.2.5 Gradient prominence prediction

In the previous section we discussed binary prominence classification only. Training a neural network that provides gradient prominence, e.g. for our approach a linear output of the neural net, has the advantage that it is not limited to two prominence classes. This gradient output can always be reduced to two or more discrete prominence classes. This facilitates comparison between binary and gradient classification. Another advantage is that only one output node is needed. This reduces the degrees of freedom substantially while the prediction of any amount of prominence is still possible. Because of the relatively low number of degrees of freedom, there is no problem in having insufficient training material, which lowers the danger of overtraining. This section deals with data selection 3 and 4 in figure 5.4. This time, however, we use a prominence scale from 0-10.
Several neural networks were trained under different conditions. However, all twelve features are always used as input. The number of hidden nodes varies from 2 to 18. Only nets with one hidden layer were used. The output of such nets is a single value around 0. Therefore the original cumulative prominence marks of the listeners (0-10) were scaled between -1 and +1 by using the formula 1/5 * prom -1. For instance prominence mark 9 yields 0.8 and prominence mark 4 yields -0.2.
As described above, the neural networks were trained with various numbers of training iterations. The number of iterations ranges from 100 up to 5500. The Training and Development test sets are biased (3 in figure 5.4) or unbiased (4 in figure 5.4). In order to present an overall performance of all the trained neural networks with varying number of training iterations, the linear correlations between the predicted prominence and the perceived prominence are calculated for the Training and the Development test set. Such a relationship may not be linear, but a higher order correlation was not tested. These linear correlation coefficients indicate the performance of the trained neural network and are used to select the optimal network. The results of neural networks giving a linear output are difficult to present. Each input feature has an output of around 0, for instance 0.4563. For rescaling the output we use the formula ((output + 1) * 5). This gives us the predicted prominence value of 7.2815 (on the original scale from 0 to 10).

5.2.5.1 Results

For the optimal unbiased trained network the highest correlation coefficient is 0.60 for the Development Test set (unbiased) and 0.64 for the Training set. This net has a topology of 12-6-1. For the unbiased condition the highest correlation coefficients are 0.70 (Development Test set) and 0.72 (Training set). The optimal net with the topology of 12-10-1 achieved these highest correlation coefficients.
In order to graphically compare the predicted prominence with the perceived prominence, figures 5.6 and 5.7 give medians (of the predicted prominence within

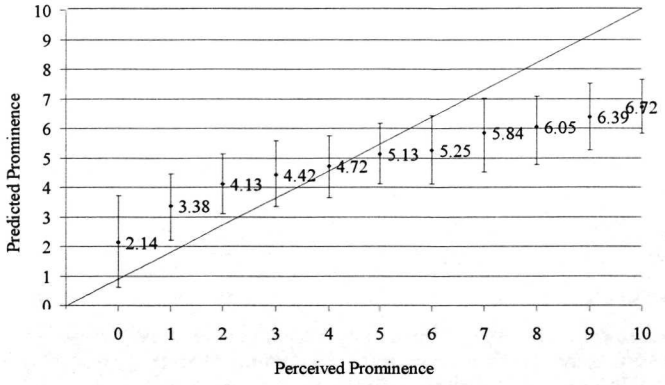


Figure 5.6: Medians ± 1 IQR (Inter Quartile Range) per median of the predicted prominence (trained with unbiased data) of the Development test set on the perceived prominence scale of 0-10. The linear correlation is r = 0.60. The dashed line gives the perfect prediction.

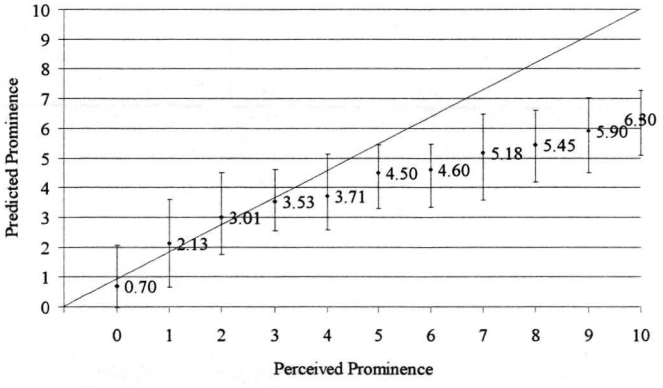


Figure 5.7: Median ± 1 IQR of the predicted prominence (trained with biased data) of the Development test on the perceived prominence scale of 0-10. The linear correlation is r = 0.70.

the prominence scale of 0-10) for the optimal neural networks for the Development test set. As error bars the Inter Quartile Ranges per median (75% and the 25% percentiles) are plotted.
The extremes of the perceived prominence scale are shifted towards the middle of the predicted prominence scale. The calculated percentiles confirm this: perceived prominence of 8, 9 and 10 is predicted with median values of 6.05, 6.39 and 6.72, respectively (figure 5.6). These values are even further shifted to the mid range for the biased training, namely median values of 5.45, 5.90 and 6.30, respectively (figure 5.7). However, for the biased training the predicted median value of 0.70 is closer to the perceived prominence 0. This is not the case for the unbiased training, where the perceived prominence of 0 is predicted to be 2.14. The error bars, expressed in IQR, show that there is a lot of overlap, indicating that the classification of each individual prominence value of 0-10 is not an easy task. In order to examine how well the predicted gradient prominence values fit into the perceived prominence scale (0-10) confusion matrices are constructed.

Tables 5.5 and 5.6 present confusion matrices of the perceived prominence versus the predicted prominence scales, achieved by the optimal neural networks (12-6-1 and 12-10-1) respectively, which have the highest correlation coefficients. In order

Table 5.5: Gradient prominence prediction versus perceived prominence. This matrix is based on the Development Test set. The prediction was achieved with unbiased data and the topology of the net was 12-8-1. Cells containing > 30 data points are boxed.

Predicted prominence \ Perceived prominence	0	1	2	3	4	5	6	7	8	9	10	Total
-2	-	-	-	-	-	-	-	-	1	-	-	1
-1	20	4	-	-	-	-	-	-	1	-	-	25
0	37	10	2	1	-	-	-	-	-	-	-	50
1	43	28	12	5	7	2	-	-	-	1	-	98
2	46	36	22	23	18	11	6	5	5	1	1	174
3	36	61	51	46	33	27	26	11	12	7	1	311
4	40	59	69	61	65	51	60	47	32	27	16	527
5	25	33	51	56	50	61	54	46	46	36	34	492
6	9	16	27	38	54	62	51	56	61	65	58	497
7	2	9	20	20	23	34	41	48	59	57	77	390
8	1	3	5	8	7	7	19	39	34	42	42	207
9	-	-	-	1	1	4	2	7	6	20	23	64
10	-	-	-	-	1	-	-	-	2	3	7	13
Total	259	259	259	259	259	259	259	259	259	259	259	2849

to put the predicted amount of prominence in such a confusion matrix, the linear output is placed on a scale of 0-10, as explained above, by rounding these values to integers. For instance the number 7.2815 is rounded off to 7. This number yields one tally in row 7 of the predicted prominence.

Especially the cells in the middle of the matrices cause confusion. The perceptually very prominent categories (8, 9, 10) are often predicted with less prominence. Training with equal numbers provides a poor non-prominence prediction; perceived prominence 0 and 1 is often predicted with 2 and 3 (see table 5.5). In addition, it is also possible that predictions are less than 0 and greater than 1. This depends exclusively on the input values. It can occur that if the neural net is fed with deviant data (for instance, because of a measurement error) the output of the neural net is also very deviant. In fact two words with a perceived prominence of 8 are predicted with a value less than 0, namely -1 and even -2. This holds for the unbiased condition. This is obviously wrong. We decided to keep these values, because it concerns only a few incidental cases.

For the training with unequal numbers (biased) the prediction on the lower part of the prominence scale is much better than for unbiased training; the cells around 0 are more filled.

Table 5.6: Gradient prominence prediction versus perceived prominence. This matrix is based on the Development Test set. The prediction was achieved with biased data and the topology of the neural network was 12-10-1. Cells containing > 10% of the total per perceived prominence scale are boxed in.

Predicted prominence \ Perceived prominence	0	1	2	3	4	5	6	7	8	9	10	Total
-5	1	-	-	-	-	-	-	-	-	-	-	1
-3	3	-	-	-	-	-	-	-	-	-	-	3
-2	20	-	-	1	1	-	-	-	-	-	-	22
-1	218	23	2	-	-	-	-	2	1	-	-	246
0	1108	106	13	12	6	8	3	1	2	-	1	1260
1	651	100	55	29	22	6	13	10	5	2	1	894
2	404	102	79	32	42	29	31	33	20	9	5	786
3	271	94	64	73	62	46	49	52	36	25	9	781
4	175	75	61	70	55	66	60	62	55	57	18	754
5	104	52	47	36	53	84	86	72	84	63	40	721
6	34	14	28	23	33	40	36	81	80	69	56	494
7	13	9	11	18	13	25	31	55	85	79	51	390
8	2	2	1	7	7	2	8	30	19	42	35	155
9	-	-	-	-	-	4	3	9	4	18	13	51
10	-	-	-	-	-	-	-	-	1	1	-	2
Total	3004	577	361	301	294	310	320	407	392	365	229	6560

5.2.6 Analyses of individual features

In chapter 4 several individual features were analyzed in general terms as prominence predictors and several histograms concerning these acoustic features were given. These histograms indicate the ability to distinguish between the two extremes of the prominence classes (namely prominence categories 0 and III). The discriminative ability of these features will be analyzed in more detail in this subsection by using neural net techniques and will be compared also with the relevant histograms as displayed in chapter 4. All four prominence categories are involved in the analyses presented in this chapter (0 and I as non-prominent, and II and III as prominent). In chapter 4 we restricted ourselves to the two extremes of the prominence categories (0 and III). Techniques such as linear discriminant analyses or CART-trees may be more powerful, but in this study simple feed-forward networks do help us to investigate acoustic correlates in detail. Our simple 1-2 neural networks give the similar results as an LDA. We chose to stay with neural networks. Individual networks with one input node and two output nodes are trained with eight of the twelve features. Features giving overall information of the sentence are omitted. Eight of the twelve features, as described in figure 5.3, were used as such single features; namely the vowel duration (1), vowel duration normalized for intrinsic vowel duration (2), vowel intensity normalized for the overall intensity of the given sentence (5), vowel intensity (sentence normalized) normalized for the intrinsic vowel intensity (6), syllable duration (7), median F_0 corrected for the median F_0 per sentence (9), range of F_0 per syllable (10), range F_0 per word (12).
Later on in this chapter we will also study some combinations of features. Individual features were analyzed by training a simple neural network with one input node and two output nodes (prominent or non-prominent). Such simple networks can be used to analyze the individual input features in two different ways. On the one hand the performance of the classification with single acoustic input features can give information about the discriminative power of these individual features, and on the other hand the neural networks themselves can be analyzed. If one uses neural networks without a hidden layer, only linear relations can be estimated, but for a preliminary examination of the data this will suffice.
As said above the design of the Training and Development Test sets are available in a biased and in an unbiased version. Consequently, a total of 16 neural networks is required that will vary only in one input feature. However, the difference between biased and unbiased training and testing remains. These neural networks were mostly trained with only 18 training iterations, which is sufficient given the few degrees of freedom. These simple nets have only 4 variables to adjust to the data.
Before presenting the performance of these 16 neural networks, the network itself will be analysed.

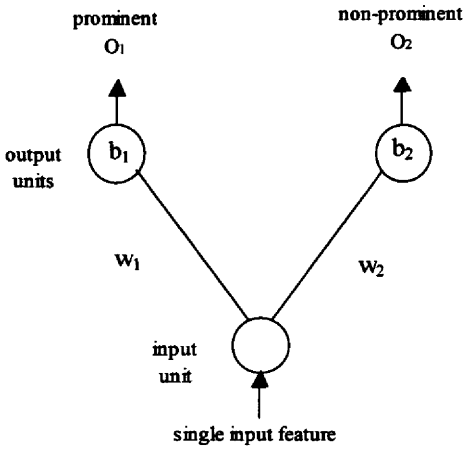


Figure 5.8: A neural network with the topology of one input node and two output nodes (1:2). Such networks are used to analyze single features and to estimate critical boundaries.

5.2.6.1 Analyzing the neural network

The neural nets used are designed in such a way that the first output node fires when the input concerns a prominent word and the second output node fires when the input concerns a non-prominent word. It can be calculated where the trained neural network puts the decision threshold in the training session. Beyond this threshold the net decides to classify this word as prominent and below this threshold value the net marks this word as being non-prominent.

The resulting activation threshold for our minimal neural network can be expressed as follows, where O_1 and O_2 are the output functions of the two output units (see figure 5.8):

$$O_1 = \frac{1}{1+e^{-(+w_1 \cdot Input - b_1)}}$$

$$O_2 = \frac{1}{1+e^{-(+w_2 \cdot Input - b_2)}}$$

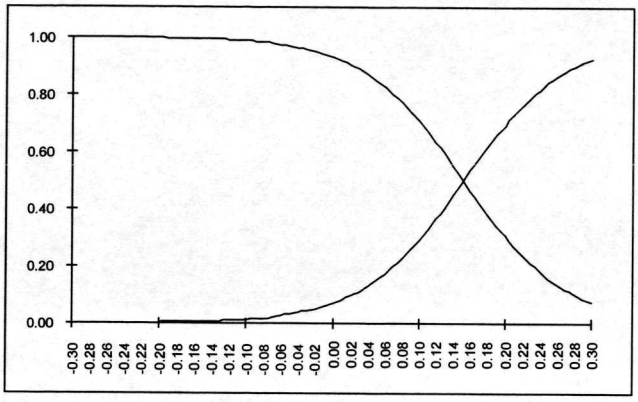


Figure 5.9: The two activation functions for the output nodes. There is one crossover point near 0.15.

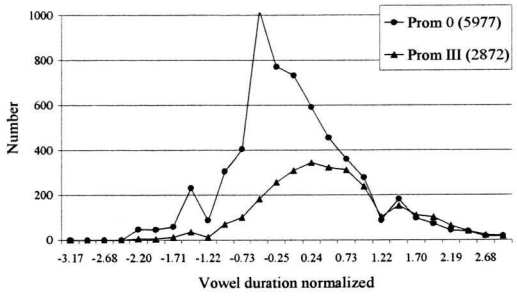
The crossover point is calculated by equalizing these two functions. This crossover point is displayed in figure 5.9. This is exactly the critical point; at this point a prominent and a non-prominent word are equally probable.

The values of these crossover points are estimated for the eight individual features. Table 5.7 gives the values for the biased and for the unbiased versions. By comparing the two columns it can be seen that the thresholds estimated by using a neural network trained with biased training input (1 in figure 5.4) are shifted to longer vowels and syllables. These vowels have also a higher intensity and show greater ranges in their F_0 movements than in the unbiased version (2 in figure 5.4). For 'vowel duration' the critical value is 0.10 s or 0.12 s, respectively. These values

Table 5.7: The estimated boundaries for 8 acoustic features. These boundaries function as a crossover point; beyond these values the neural network classifies the given data as belonging to a prominent word.

Feature	Estimated boundaries	
	Unbiased	Biased
Vowel duration (s)	0.10	0.12
Vowel duration normalized (z-score)	0.09	0.75
Vowel normalized for sentences intensity (dB)	77.89	79.69
Vowel intensity (dB) normalized (z-score)	0.17	0.74
Syllable duration (s)	0.21	0.24
Range F_0 per syllable (st)	3.36	4.19
Median F_0 per syllable (st above the sentence median F_0)	0.23	4.33
Range F_0 per word (st)	3.75	4.69

a) Biased:



b) Unbiased:

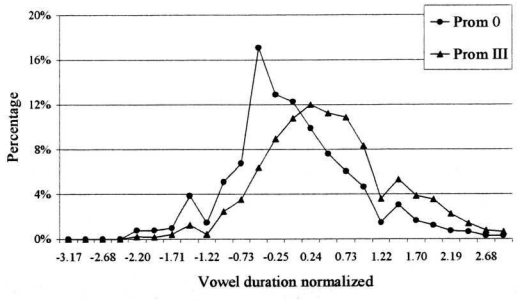


Figure 5.10: The biased and unbiased distribution of the vowel duration normalized for intrinsic vowel duration. The peaks at the edges of the histograms are due to the bin distribution. (More details are explained in section 4.2.2.3.)

indicate that vowels longer than 0.10 s or 0.12 s are classified as prominent. This unbiased threshold of 0.10 s is similar to the crossover point in figure 4.9, which actually represents unbiased data, because in these figures percentages are presented. The threshold does not have to be identical because the data are not identical (in the histogram only prominence class 0 and III are presented, whereas the nets are trained with 0 and I as 'non-prominent' and II and III as 'prominent'). The unbiased threshold (0.09 s) of 'vowel duration normalized for intrinsic vowel duration' shows that, if a vowel is classified as prominent (prominence class II and III) it must be longer than the threshold value of the class it belongs to. If the normalized duration had been exactly equal to the mean of its class, this value would have been 0. The network trained with biased input data of the vowel duration corrected for the intrinsic vowel duration (vowel duration normalized) places the threshold at 0.75 (z-score). This is a large shift to the right. This can be explained by looking at the distribution of this feature as shown in figure 5.10. Graph a) presents the biased data of the two extremes of the prominent class (0 and III), and graph b) presents percentages, which are corrected for the differences in numbers in the two extremes of the two prominence classes. Prominence class III lies below the prominence class 0 histogram as shown in graph a). The neural network places the threshold far to the

right so that chance distribution of the frequency of occurrence determines the classification.
A similar situation occurs for 'vowel intensity corrected for intrinsic vowel intensity' and for 'median F_0 per syllable'. These three features (vowel duration normalized for intrinsic vowel duration, vowel intensity (dB) normalized for intrinsic vowel intensity, and the median F_0 per syllable) are probably not useful in a biased classification situation. The classification results, presented later in this section, will show whether these three features are useful for unbiased classification situations, e.g. for disambiguating the meaning of sentences. The thresholds of the other features related to the vowel intensity are as expected. The 'biased' thresholds have further shifted towards greater intensity in order to cope with the greater probability that data belong to non-prominent words (60% of the data belong to non-prominent words, whereas only 40% belong to prominent words). The F_0 range thresholds measured on syllables (3.36 st or 4.19 st) and on words (3.75 st or 4.69 st), respectively, show that the F_0 range must be larger for the word condition than for the syllable condition. The thresholds belonging to the unbiased version correspond to the crossover points of the histograms as in figure 4.6.

5.2.6.2 Analyzing the performance of the individual features

The performances of these simple neural networks are of interest because a comparison can be made with the amount of overlap of the histograms as displayed in chapter 4 and as expressed in table 4.3 in that chapter. In this comparison, not only the absolute recognition rate is important but also whether a ranking can be made in order to work out the most useful features for prominence classification. A combination of the 'best' features brings us closer to the training of the 'optimal' neural network for ultimate prominence classification for a binary prominent / non-prominent distinction as well as a gradient prominence classification.
Table 5.8 presents the recognition rates for the above-discussed simple neural networks with single input features for the biased and unbiased versions of training and testing. The recognition rates are only given for the Development Test set, and are separately presented for prominent and non-prominent input. The chance level of the frequency of occurrence is 50% for the unbiased recognition rates, whereas for the biased version this level is about 60% according to the prominent / non-prominent distribution as explained above. As expected from the estimation of the thresholds, 'vowel duration normalized for intrinsic duration', and 'vowel intensity normalized for the intrinsic intensity', and 'median F_0 per syllable', performed slightly above chance level for the biased trained neural networks. Because of the differences in the biased distribution, the non-prominent recognition in biased condition is always better than the prominent recognition. In the case of the unbiased condition, recognizing prominent words is better when using intensity features; the other features do not show such a constant preference.

Table 5.8: The performance of the individual acoustic features trained with a simple neural network (1-2) without a hidden layer. The overall recognition rate on the Development test set for the biased and the unbiased data is given together with the separate results for the prominent and non-prominent recognition. Also the ranking of the overall scores is given together with the ranking based on the histogram overlap as presented in chapter 4, section 4.3, table 4.3.

		Unbiased	Ranking	Biased	Ranking	Ranking overlap
Vowel duration	All	**65.97%**	5	**66.01%**	4	6
	Non-prom	65.87%		81.05%		
	Prom	66.07%		43.88%		
Vowel duration normalized	All	**61.01%**	8	**61.57%**	7	8
	Non-prom	62.23%		83.33%		
	Prom	59.78%		29.57%		
Vowel intensity corrected overall	All	**66.86%**	4	**65.67%**	5	4
	Non-prom	59.86%		74.16%		
	Prom	73.87%		53.18%		
Vowel intensity normalized	All	**62.11%**	7	**61.63%**	6	7
	Non-prom	56.69%		77.95%		
	Prom	67.54%		37.63%		
Syllable duration	All	**71.34%**	2	**70.14%**	1	3
	Non-prom	71.18%		79.82%		
	Prom	71.50%		55.89%		
Range F_0 syllable	All	**69.02%**	3	**67.73%**	3	2
	Non-prom	71.38%		77.18%		
	Prom	66.67%		53.82%		
Median F_0 corrected	All	**64.45%**	6	**60.11%**	7	5
	Non-prom	65.95%		90.14%		
	Prom	62.95%		15.93%		
Range F_0 word	All	**71.85%**	1	**69.83%**	2	2
	Non-prom	72.80%		77.26%		
	Prom	70.90%		58.91%		

Similar to the ranking based on the overlap of the histograms in table 4.3, a discriminability ranking can also be derived from the recognition rates of the neural networks. In table 5.8 such a ranking is displayed for the biased and the unbiased condition. Rankings 2 and 1 are interchanged for the unbiased and biased condition, indicating that syllable duration and F_0 range per word are the features with the highest discriminatory power in both conditions. The F_0 range per syllable follows

these two features. Places 4 and 5 are interchanged as well; this means that the unnormalized version of vowel duration and intensity (sentence normalized) are interchanged. The ranking for the histogram overlap (see section 4.3 table 4.3) follows more or less the network ranking. Place 2 is interchanged with place 3. As already indicated in chapter 4, duration and intensity normalization does not further improve prominence classification. The percentages of correct recognition are poorer for normalized features than for the unnormalized counterparts (see table 5.8).

5.2.6.3 Summary and conclusion

The ranking of the recognition results differs somewhat for the biased and unbiased version, as well as in comparison to the ranking of the amount of overlap (see chapter 4), but the trend is clear. F_0 ranges per word and syllable duration are useful features. Vowel intensity and vowel duration are also useful, while their normalized counterparts do not improve the classification of this speech material.
In the next section it will be estimated whether certain combinations of features will further improve the classification task.

5.2.7 Analyzing combinations of features

Only a few paired feature combinations were selected to train a neural network and, for simplicity reasons, only the results of the unbiased trained nets are presented here. Preferably, only combinations should be involved which do not intercorrelate highly (see table 5.1), ensuring that each feature with the combination contributes independent information.
The recognition results of the various combinations are presented in table 5.9. The

Table 5.9: The correct recognition rates (% correct) of a number of acoustic feature combinations, expressed in overall correct recognition rates on the Development Test set for the unbiased data.

Feature combination	Dev. Test (%)
Range F_0 word - syllable duration	75.50
Range F_0 word - vowel intensity	74.48
Range F_0 word - vowel duration	74.04
Syllable duration - vowel intensity	72.69
Syllable duration - sentence speaking rate	71.45
Vowel duration - vowel intensity	69.33
Vowel intensity - overall intensity	65.98
Vowel duration - sentence speaking rate	66.01
Vowel intensity normalized - overall intensity	62.77
Vowel duration normalized - sentence speaking rate	61.15

results are ordered according to recognition rates. It is striking that the range of F_0 per word, in combination with the syllable duration, vowel intensity, or vowel duration gives the best classification. The acoustic feature combination of syllable duration and vowel intensity also gives acceptable results.
Although the sentence speaking rate yields some effect (chapter 4, figure 4.13), this additional information did not provide an improved performance. The feature of sentence speaking rate in combination with either vowel duration or syllable information gives a performance of 66.01% or 71.45%, respectively. However, the single features analyses gives 65.97% and 71.34% performance. In addition, the combination of sentence speaking rate and vowel duration normalized for intrinsic vowel duration does not improve performance (61.15% with sentence speaking rate versus 61.01% for the single feature). The trained networks with two-wise feature combinations indicate that the unnormalized features are the most promising ones. For this kind of speech material, normalizations do not increase the prominence classification substantially. Normalization cannot (yet) be implemented in such a way that it increases the performance.

5.2.8 Prominence classification with an 'optimal' feature combination

Based on these results we combined F_0 range per word, syllable duration, vowel duration and vowel intensity as a promising feature set.
A neural network trained with these four 'basic' features performs almost as well as

Table 5.10: Performance on the Development Test set of several neural networks with 4 'basic' acoustic input features.

		Unbiased training			Biased training		
Num of hidden nodes		Non-prom	Prom	%	Non-prom	Prom	%
4	All	Dev. test		**78.11**	Dev. test		**78.02**
	Non-prom	1871	655	74.07	3145	760	80.54
	Prom	451	2075	82.15	682	1973	74.31
6	All	Dev. test		**77.95**	Dev. test		[illegible]
	Non-prom	1863	663	73.75	3163	742	81.00
	Prom	451	2075	82.15	682	1973	74.31
8	All	Dev. test		**78.35**	Dev. test		**78.26**
	Non-prom	1871	655	74.07	3169	736	81.15
	Prom	439	2087	82.62	690	1965	74.01
10	All	Dev. test		**78.11**	Dev. test		**77.94**
	Non-prom	1874	652	74.19	3128	777	80.10
	Prom	454	2072	82.03	670	1985	74.76
12	All	Dev. test		**78.48**	Dev. test		**78.08**
	Non-prom	1869	657	73.99	3150	755	80.67
	Prom	430	2096	82.98	683	1972	74.27
14	All	Dev. test		[illegible]	Dev. test		**77.96**
	Non-prom	1900	626	75.22	3153	752	80.74
	Prom	433	2093	82.86	694	1961	73.86

a neural network trained with all twelve features. The neural network trained with the four features shows an overall performance of 79.04% correct prominence and non-prominence classification on the Development test set under the unbiased condition (see table 5.10). The best biased condition gives a performance of 78.29% correct. The performance of a net trained with all twelve features still performs somewhat better though (unbiased condition 82.01%, biased condition 80.11%, see table 5.2) on the Development test set.

5.3 Discussion and conclusion

Twelve acoustic input features for binary prominence prediction yield 82% (unbiased condition) and 80% correct (biased condition) classification on the Development test set and 79% and 78% correct classification on the Independent test set. Higher results are found in the literature. Kießling (1996) achieved 82.8% (spontaneous speech material) and 95% (read aloud speech material, simple sentences) correct classification (for accent / non-accent in their terminology), using a net with 276 input features including textual information. Such a comparison is not completely fair, because we aim at different goals. The statistical or brute force method used in Kießling (1996) aimed at high recognition rates whereas our approach aimed different perspectives of acoustics and classification. Our prominence classification results in this chapter were achieved using solely twelve acoustic input features.

The binary prominence predictions are as consistent as listeners are. On an Independent Test set (unbiased data) these nets achieve 77% correct ($\kappa = 0.56$) with the most optimal listener (see also chapter 2). The agreement is even higher for the biased data: namely $\kappa = 0.57$. These agreements do not differ from the agreement between listeners which is on average $\kappa = 0.50$. Thus our neural networks are indistinguishable from naive listeners for assigning prominence.

The attempt to predict gradient prominence is much more complicated than binary prominence prediction. A correlation of $r = 0.60$ (unbiased condition) and $r = 0.70$ (biased condition) is achieved. In principle high correlations could only be a indication of high recognition results. Looking at the underlying confusion matrices it appears that the middle range of the prominence scale (0-10) is an area of confusion, and that the extreme of 10 (very prominent) is rarely predicted. It may be that our design of perceptual prominence judgments is not constructed to allow for a really accurate prediction of a gradient scale.

The analyses of the individual acoustic features confirm that the four 'basic' acoustic features, namely vowel duration and intensity, syllable duration and F_0 range per word, yield performances of 79% (unbiased condition) and 78% (biased condition) correct prominence classification on the Development test set. Despite the fact that normalizations provide no further improvement when used as two-wise features, using all twelve input features still shows somewhat better performance

(table 5.2). Exactly what the effect is of the eight remaining features still needs detailed investigation. Also, it can be concluded that the use of a hidden layer provides more accurate prominence classification (table 5.2), also in comparison with an LDA. This means that there is no simple linear relationship between prominence and acoustic features. Therefore, the linear representation in the histograms of chapter 4 is a first approximation only.