

PITCH, LOUDNESS, AND SEGMENTAL DURATION CORRELATES: TOWARDS A MODEL FOR THE PHONETIC ASPECTS OF FINNISH PROSODY

*Martti Vainio*¹
*Toomas Altonsaar*²

¹Department of Phonetics, University of Helsinki, Finland

²Acoustics Laboratory, Helsinki University of Technology, Finland

ABSTRACT

Neural networks are used widely today for modeling a variety of different aspects of spoken language. We use them to model Finnish lexical prosody with an aim to shed light on the interaction between the main prosodic parameters: segmental durations, loudness and pitch. We have analyzed the performance of a group of networks which were all trained to generate values for a different prosodic parameter given similar input information. The experiments were performed on speech material contained within our Finnish speech database.

1. INTRODUCTION

No comprehensive model has been developed for Finnish prosody and the previously conducted studies lack uniformity which inhibits the extraction of explicit knowledge upon which functional models could be based. Our study serves as a starting point for research which aims to remove this deficiency. At this point we are striving to determine the correlates of pitch, loudness, and segmental durations for Finnish prosody on the lexical level. We have developed methods for modeling prosodic information with neural networks and our aim is to form a model which integrates these three factors. Our model should improve the quality of speech synthesis and recognition as well as increase the general knowledge of Finnish phonetics.

Since we use neural networks we do not necessarily expect to infer a number of explicit rules like those found in traditional rule-based speech synthesis. Regardless of this, neural networks are not intended to be simply a substitute for explicit rules. Instead, they can be viewed as an aid in finding general patterns in speech material and to support the search for any such explicit rules that may be governed by some, possibly, higher level structural components.

2. NEURAL NETWORK ORGANIZATION AND DATA REPRESENTATION

The task for the networks is to compute a value (duration, average loudness-level, or average pitch) for a known phoneme context defined by the phoneme sequence environment and the word context. The problem at hand is to find a network organization, a data representation, and a way for training the network successfully from real speech data. We have so far experimented with multi-layered feedforward networks that estimate either segmental durations, loudness, or pitch values independently as

well as networks which estimate all three factors concurrently. The general structure of the networks and the input vector coding scheme can be seen in Figure 1.

The natural form of a neural network output is a real number. Since the distribution of duration follows a logarithmic, rather than a linear time scale, it is well motivated to train the network to yield the logarithm of the duration value. For loudness and pitch a simple linear mapping is sufficient.

The network's hidden and output nodes use a sigmoid function and the error feedback term is simply (target - output) instead of the usual; $output * (1 - output) * (target - output)$.

The original time functions for loudness and pitch have to be reduced to a single value for each segment. For this we chose the central 1/3 part of a segment to define the averaging span.

The input data for the network is a sequence of phonemes. There are many ways in which the phoneme sequence can be represented. We used a data representation which assumes that a phoneme has certain intrinsic values (e.g., inherent duration) that are unique to itself and that the same phoneme has some class intrinsic values that it shares with other phonemes (e.g., low vowels have higher inherent loudness levels). This allows for our phonemes to exist in a hierarchical class system. E.g., phoneme /i/ belongs to the class *front-vowel* as well as to the class *vowel*. Furthermore, we assumed that the phonemes influence each other when in close proximity and that the further the separation between the segments the more general the influencing factor is. E.g., in Finnish the length of a primary stressed vowel influences the duration of the following secondary stressed vowel [2].

We therefore adopted a coding scheme which takes all of these assumptions into account. The estimated phoneme is coded according to its identity, class and length (three real numbers as in [5]); its nearest neighbors are coded according to their classes and length (two real numbers); for any further neighbors we code its length and the vowel-consonant distinction (two real numbers). This forms a multi-resolution representation of the input phoneme string.

Each training vector also includes a real number for the length of the word and the estimated phoneme's position in that word. With the varying window width in our tests, the vector size varied from 2 to 25 numbers. We also experimented with 2 to 16 hidden nodes to determine the optimum network size for a problem.

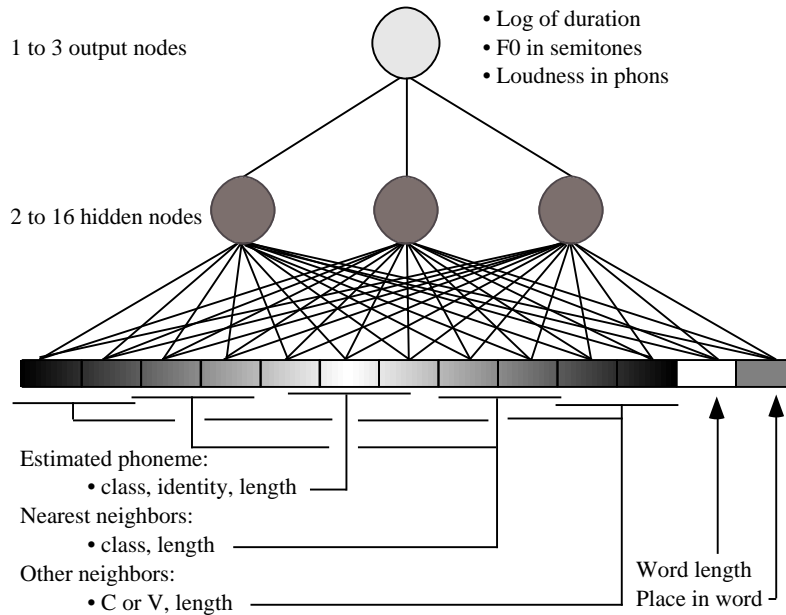


Figure 1. Neural network organization and multi-resolution data representation in word level segmental duration, loudness and pitch studies.

Since we were looking for correlations between the different prosodic factors we decided to give each type of network an identical training vector, i.e., all networks were presented with equivalent information.

3. EXPERIMENTS AND RESULTS

All of the experiments were carried out with our Finnish speech database of isolated words. The database is integrated on top of the QuickSig signal processing environment which runs under Lisp/CLOS [1]. 2000 phonetically balanced words from a single speaker were used for training (2/3) and evaluation (1/3). The use of a single speaker was due to dialect differences between the speakers in the database. Our experiments did yield similar results for the other speakers with only the average error rate varying.

We were mainly interested in whether the length of the context window had any effect on the network's ability to estimate the target values. We used different methods for analyzing the output errors. One method can be seen in figures 3-5. They each present the performance of 125 averaged networks as a function of the input vector's composition. The x-axis represents the length of the right context (in terms of phonemes) and the y-axis the left context, respectively. Each grid point shows the average error of five optimal networks each trained for 500 training set passes. The results are presented in a way that produces a spectrogram-like image.

Another way of presenting the network errors can be seen in figure 2. It shows an error decomposition for a single network trained to estimate vowel durations. The estimates are

compared with actual words in the database not used in training. The error decomposition window is mouse-sensitive and each individual value can be inspected and the comparison word can be listened to or viewed in a separate window. This method has actually been useful in locating faulty segmentations in the database – the largest errors usually point to labeling errors, or to words with foreign origin, or simply to words which does not share structure with other words in the database.

3.1. Duration Networks

Finnish is a so called quantity language with two distinctive lengths for segmental durations. The quantity of each segment has to be distinctive in all circumstances. This makes the system fairly sensitive and places an extra burden on any model used to estimate or generate segmental durations.

The duration networks performed generally better when they were trained to estimate values for a subclass of phonemes, i.e., when they were more specialized. For optimal performance we trained 16 networks which were categorized according to natural phonetic classes. The average error count varied from 5% for long stops to 14.2% for short fricatives and 21.1% for semivowels and liquids. The average error for the entire set of networks was about 12.5%, which is below the difference limen (20%) for segmental durations as reported by Klatt [3] but above the 5% threshold reported by Eek [4]. A study by Karjalainen et al. [5] proposed a value of 20% for Finnish.

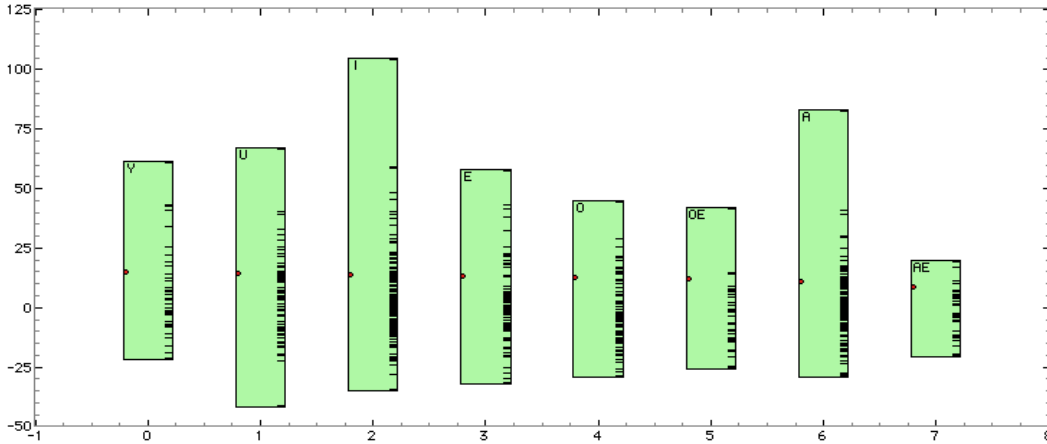


Figure 2. Error decomposition for a network trained to estimate duration values for short vowels. The relative errors (in percent) for the vowel tokens are indicated by small lines on the right side of the bars. The small dot on the left side of each bar shows the average absolute relative error for each vowel. The y-axis is the relative error in percent. The bars are ordered according to decreasing average absolute relative error.

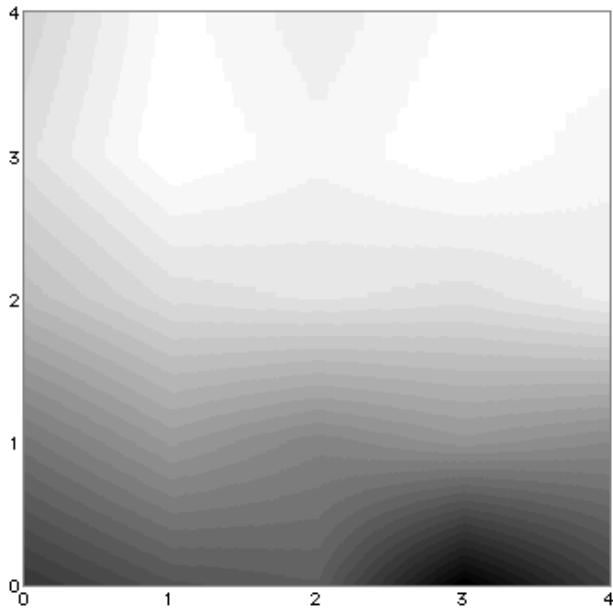


Figure 3. Average absolute relative errors for duration networks as a function of the input vector's composition. The x-axis represents the right context and the y-axis the left context. Each pair is an average result of five separate neural networks trained for 500 training iterations of the training set. Lighter shaded areas represent smaller average error; in this case the (left=3, right=1) window yielded the best result (17.0%). The difference between the minimum and the maximum in error (found at (0,3)) is approximately 5.3%.

3.2. Loudness and Pitch Networks

The absolute loudness values in our database varied considerably and had to be normalized. We adopted the following normalization scheme: first a maximum in the loudness-level signal for an individual word was located, then the phone with the maximum peak value was identified and a bias was added to the whole signal. The bias was simply an average loudness-level value derived from the entire database for each phoneme. E.g., if the maximum peak occurred during vowel [i], a difference between the average [i] value and the reference value (taken from the inherently loudest vowel [a]) was subtracted from the signal. Unlike duration networks, loudness and pitch networks required no specialization and one net for each parameter was sufficient. For the loudness network the average error was ≈ 2.2 phon and for the pitch networks $\approx 3.5\%$ on the Hz scale.

Average Error %	Context left, right	Input Nodes	Information in Input Vector
28.8	0,0	1	position in word
27.1	0,0	1	identity
26.0	0,0	1	length
22.4	0,0	1	length + identity*
22.2	0,0	2	length + identity
20.4	0,0	3	length + identity + class
15.2	2,1	14	no word length given
13.8	2,1	14	no pos. in word given

Table 1. The effect of increasing information content in the input vectors for duration networks. *Coded as one number.

Table 1 presents the effect of increasing the amount of information presented to a network. For this test 8 separate networks were trained on a subset of material used for figure 3. When only the position of the phoneme in the word was used the average absolute relative error was 28.8%. When the length of the phoneme as well as its identity was coded as two numbers the error dropped to 22.2%. Adding contextual information from the neighboring phonemes, i.e., two neighboring left phonemes and one neighboring right phoneme, reduced the error to 13.8%.

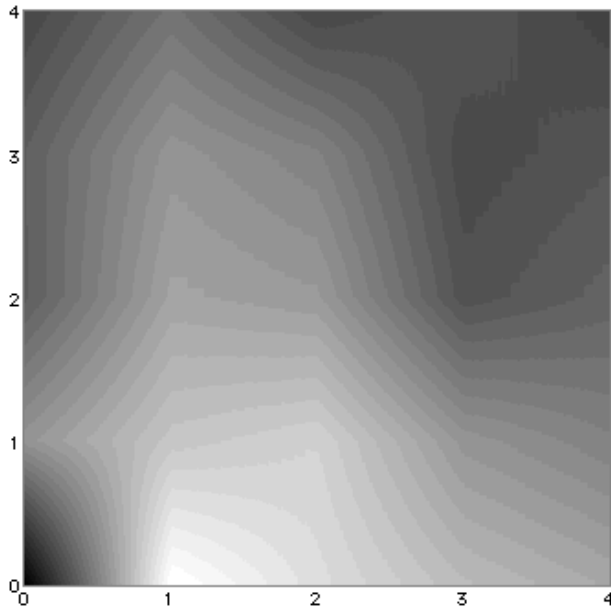


Figure 4. Average absolute errors for loudness networks as a function of the input vector's composition, voiced segments only. Maximum error was 2.4 phon at (0,0), minimum error was 2.2 phon at (0,1).

4. DISCUSSION

Psychoacoustically our results indicate that the networks model the complex prosodic parameters in an adequate manner. Estimating duration proved to be the most difficult task and it had to be divided into subproblems for the average error to decrease below 20%, the difference limen suggested in [3]. Networks modeling loudness achieved an average error of 2.2 phon (1 phon is generally considered just noticeable). Modeling pitch was achieved with an error of 3.5%. This amounts to about 0.6 semitones at 100 Hz and is well below the 1.5 to 2 semitone threshold for speech [6].

The distance between the minimum and the maximum among the distributions in figure 4 is only 0.2 phon for the loudness networks and only 0.3% in figure 5 for the pitch networks. This seems to suggest that the phoneme context has very little effect on the loudness and pitch levels of a single phoneme segment. This contrasts with the segmental durations which seem to be much more dependent on their environment (figure 3). Concurrent

networks that had three output nodes, one for each parameter, were also trained. However, their performance level was always worse than networks that were specialized for each prosodic attribute.

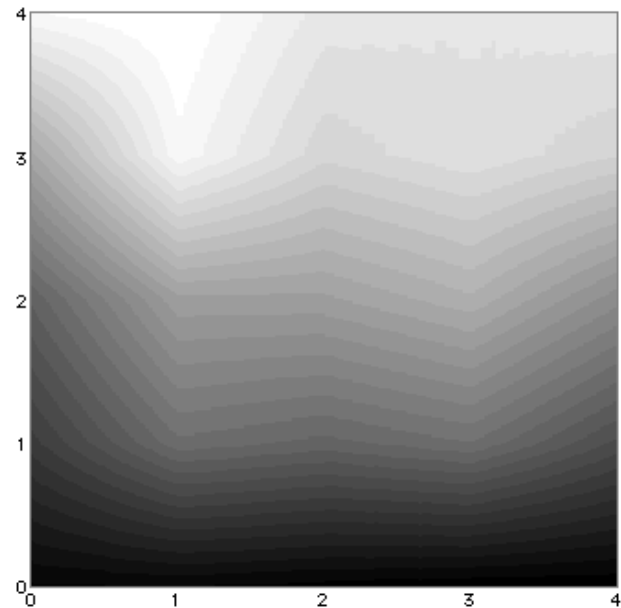


Figure 5. Average absolute relative errors for pitch networks as a function of the input vector's composition, voiced segments only. Maximum error was 3.8% at (0,0), minimum error was 3.5% at (4,1).

5. ACKNOWLEDGMENTS

We would like to thank professor Matti Karjalainen who gave us valuable direction during this study. The project has been financed by the Academy of Finland.

REFERENCES

1. Karjalainen, M. and Altosaar, T., "An Object-Oriented Database for Speech Processing," *Eurospeech-93*, Berlin, 1993.
2. Wiik, K., *Finnish and English Vowels*, Annales Universitatis Turkuensis, Series B, Tom 94, Turku, 1965.
3. Klatt, D. H., "Linguistic Use of Segmental Durations in English," *JASA* 59: 1208-1221, 1976.
4. Eek, A., "Just-Noticeable Differences of Duration for Some Word Structures: II," *Estonian Papers in Phonetics*, pp. 21-26, Tallinn, 1978.
5. Karjalainen, M. and Altosaar, T., "Phoneme Duration Rules for Speech Synthesis by Neural Networks," *Eurospeech-91*, Genoa, 1991.
6. 't Hart, J., Collier, R., and Cohen, A., "A perceptual study of intonation," Cambridge University Press, Cambridge, 1990.