MODELING THE MICROPROSODY OF PITCH AND LOUDNESS FOR SPEECH SYNTHESIS WITH NEURAL NETWORKS

Martti Vainio[1]
Toomas Altosaar[2]

[1]University of Helsinki, Department of Phonetics
P. O. Box 35 (Vironkatu 1 B), FIN-00014 HY, Finland, http://www.helsinki.fi/hum/hyfl/
[2]Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing,
Otakaari 5A, P.O. Box 3000, FIN-02015 HUT, Finland, http://www.acoustics.hut.fi/

ABSTRACT

In this study of Finnish microprosody, two prosodic parameters — pitch and loudness — were modeled with artificial neural networks. The networks are of the general feed forward type trained with backpropagation. For each phoneme, the network predicts a series of either pitch or loudness values on the basis of information of the phoneme's phonologically motivated features and its phonetic environment. The tests we have run so far indicate that the neural networks are highly successful and accurate in modeling the micro-level behavior of both pitch and loudness. The tests were conducted on isolated word material but some preliminary results obtained from sentence material are also discussed.

1. INTRODUCTION

Pitch-related microprosodic variation has been well attested for several languages including Finnish. For instance, the fundamental frequency difference between open and close vowels and the effect of immediate consonant context on the F_0 of a vowel seem to be universal [10], [1], [9]. Similar variation can be observed with regard to loudness. The most well known phenomenon is the difference between the inherent loudness levels of, e.g., open vs. close vowels and sonorant vs. obstruent consonants [5].

The microprosodic characteristics can be seen as the lowest level of a multi-layered prosodic system producing the final suprasegmental realization of speech. They are not generally seen as a part of the linguistic-prosodic pattern of the utterance, but rather to be segmentally conditioned. That is, they reflect the gestures necessary for producing the specific articulatory movements for various vowels and consonants.

In speech synthesis, microprosodic modeling has usually been fairly scarce — the developers have concentrated on more salient and urgent problems and the modeling has usually remained on a first approximation level. In general, speech synthesizers use some information about the intrinsic pitch, loudness and duration of speech sounds which are changed algorithmically according to certain rules that take the sounds' context into account. The microscopic changes within the time-varying parameters of the sounds have not been paid much attention to, although most synthesis systems do model the timing of F_0peaks and differences in F_0slopes and onsets after different consonants. It is probable that the inclusion of microprosodic variation would improve the naturalness and even the intelligibility of synthesized speech.

It can be argued that microprosodic variation is analogous to variation in other aspects of speech in that there are both phenomena that are extremely common and phenomena that are extremely rare. The rich combinatorics of natural language makes the number of possible combinations of units very large. Consequently, the individual phenomena that are rare in themselves become common when seen as a group and occur frequently in running speech or text. This makes it practically impossible to gather databases that can serve as a training basis for all the phenomena and combinations in speech (even in some constrained domain, such as microprosody). This calls for models that can make generalizations of some kind and generate accurate predictions for patterns that are absent in the database.

Neural networks are known for their ability to generalize according to the similarity of their input but at the same time known for being able to distinguish different outputs from input patterns that are superficially similar. This means that the networks can learn to predict patterns it has never seen before — a fact that makes them an ideal candidate for building models from imperfect data for the highly complex phenomena that prosody comprises in all its levels.

The network architecture used here, as well as the data representations for both types of networks, was the same throughout the tests since the problem at hand is quite similar — to model microscopic variations in two time-varying parameters that occur in similar circumstances and are for the most part governed by the same factors.

The models were trained speaker-dependently, i.e., one or more models for each speaker were generated. The study was carried out on the object-oriented QuickSig signal processing environment, which is programmed in LISP/CLOS [4]

2. TRAINING AND EVALUATION DATA

The tests presented here were conducted on a database of about 2000 hand-labeled isolated words spoken by two male Finnish speakers. The words in the set include most bi-phonemic sequences found in Finnish and some interesting tri-phonemic sequences (mostly consonant clusters). The words were further divided into two training and two evaluation sets with a ratio of 2 to 1, respectively.

We used nine points (or frames) for the relative linear position of the estimated value within the phoneme. Thus, each phoneme in the set produced nine training elements for the networks. The total number of training elements varied from about 500 to 20000.

2.1. Input Data Normalization

The signal amplitudes in our database are not homogeneous due to slightly different conditions during the recording phase — the distance between the speaker's mouth and the microphone, for instance, could not be kept totally fixed. For this reason we had to implement a normalization scheme to keep the inputs for the loudness networks as constant as possible. Our scheme is as follows: first a sonority table is calculated for each phone/phoneme in the database for each speaker (this table corresponded with the ones reported in the literature with the open vowels being the loudest, followed by mid and close vowels [5]). Second, each loudness signal is shifted according to the peak (which invariably falls on the first syllable nucleus) and the vowel in which the peak occurs. For instance, if the peak occurs in the vowel [*a*] (the loudest one), the signal is shifted so that the peak value becomes 100 phon — if the peak occurs in some other vowel, the signal is shifted in such a way that the peak value will be 100 phon minus the value in the sonority table. Thus, e.g., a peak occurring in [i] will result in a value of $100 - 4.8 = 95.2$ phon. This is obviously not the best way to normalize the loudness signals but it had a positive effect on the networks' performance.[1] See section 4 for details on the normalization of the F_0-curves.

3. NEURAL NETWORK ORGANIZATION AND INPUT CODING

The neural networks used in this study are of the general feed forward type trained with backpropagation. The networks consist of three layers — input, output and a hidden layer. The output layer consists of one node which outputs either a fundamental frequency value in (coded) semitone (later converted to an absolute Hertz value) or a loudness value in (coded) phon. The input has eighteen values for a distributed coding scheme (see below). The hidden layer has eleven nodes — the optimal number was determined empirically. Figure 1 shows the network's architecture as well as its input coding strategy.

The input coding follows a distributed scheme used successfully in our earlier studies of Finnish lexical prosody [6], [7] and [8]; this was an adaptation of the scheme used by Karjalainen and Altosaar for predicting segmental durations in Finnish [3]. A sequence of phonemes is represented by a set of linguistically motivated features that are straightforward to calculate from a string of phonemes and require no structural analysis of the input text[2]. The features are: phoneme identity (e.g., */a/*), phoneme class (e.g., nasal), phoneme broad class (consonant vs. vowel) and quantity degree (short vs. long). Each input vector also includes three values representing the information about the context for the estimated value or frame. These are: length of the word (as the number of phonemes in the word), position of the estimated phoneme in the word and the position of the estimated frame in the phoneme — the estimation for each phoneme thus consists of nine equidistant frames or points within the span of the phoneme. The input vector covers a seven-phoneme window by providing information about three phonemes on both sides of the estimated one. Moreover, the context is coded in a manner which provides more resolution (i.e., more detailed information) for the nearby neighbors and less resolution for the further neighbors. Each input value is coded as a real number between zero and one. See Figure 1 for more detail.

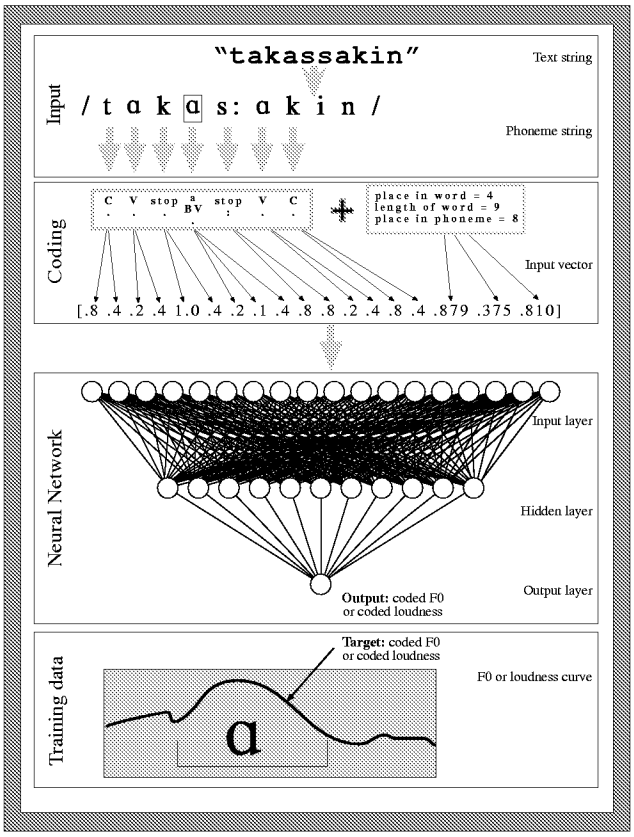

Figure 1: The neural network input, coding and architecture. The example shows the coding for the vowel / in the word *takassakin* ('in the fire-place, too'). A seven-phoneme window is used; the three features for the vowel are phoneme identity (*a* = */a/*), its class (BV = back vowel) and its length (. = short). The additional information in the training vector includes: the estimated phoneme's place in the word, the length of the word and the estimated frame's position in the phoneme.

4. RESULTS

The performance of both types of networks is summarized in Table 1. The results for loudness are somewhat preliminary since the networks were trained on data that was normalized by according to maxima within words; i.e., the network estimates, not only the contour within the phone, but within the whole word as well.

Figure 3 shows a comparison of the actual fundamental frequency values and the neural network estimates for six different cases. See the caption for more detail. Somewhat similar cases for loudness can be seen in Figure 4.

[1] The normalization scheme does not take into account the differences in the stress level between the words. However, this does not seem to be a problem, for the words in the database were articulated in a very monotonous and neutral manner.

[2] The only structural analysis we have experimented with so far has been the syllabification of the input text. This, however, had very little positive effect on the networks' prediction capability [7].

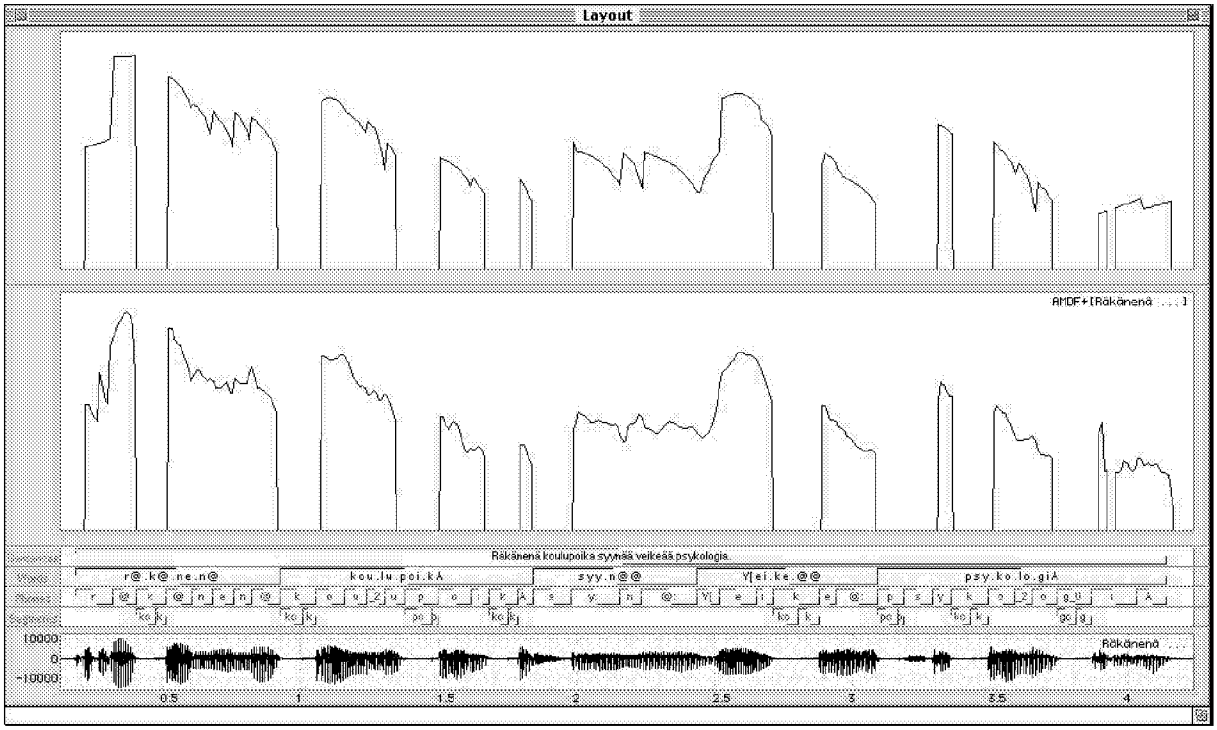

Figure 2: The neural network estimate for pitch (upper pane) and the actual F_0-contour (below). The estimates for each phone were shifted according to the average F_0 of each segment in the original pitch segment. Although the network was trained on vowel data only, the estimates for other voiced phones are also shown.

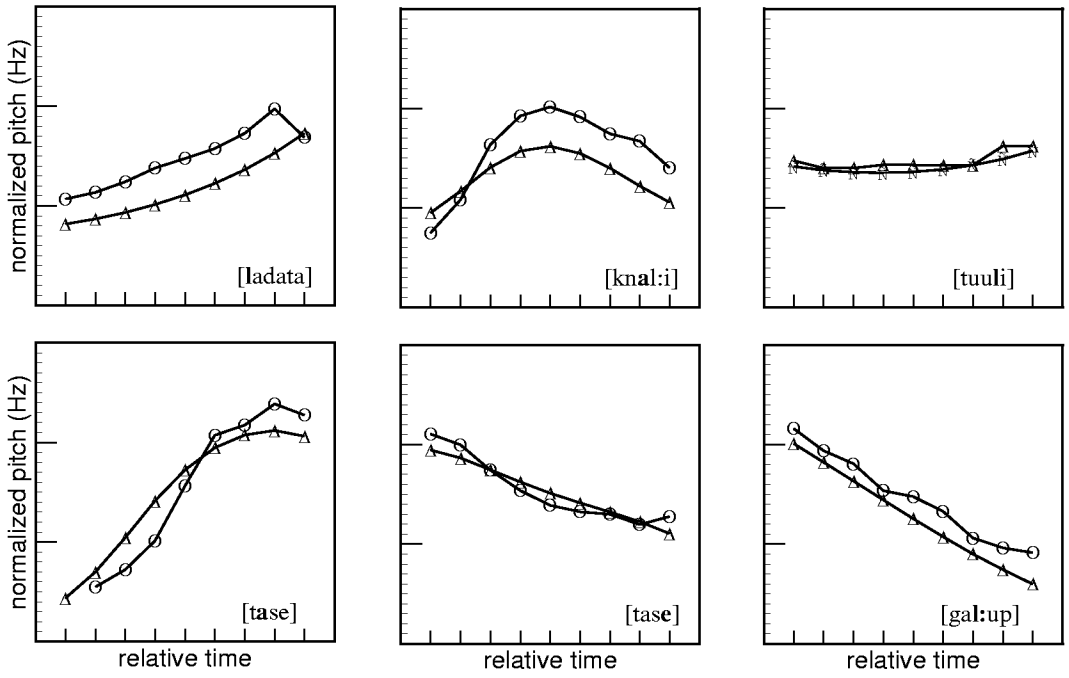

Figure 3: Estimated pitch and actual values for [*a*] in the words *knalli* and *tase*, for [e] in *tase* and [l] in *ladata*, *gallup* and *tuuli*. The vowels are estimated with a network that was trained on all voiced phones; the l-estimates represent a specialized network trained only on [l] phones. The triangles represent neural network estimates and the circles the actual F_0 values. The x-axis represents the nine estimation frames for each phone.

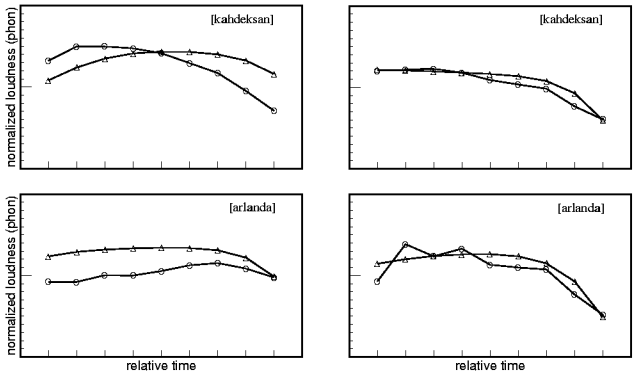

Figure 4: Estimated loudness and actual values for [*a*] in the words *kahdeksan* and *arlanda*. The triangles represent neural network estimates and the circles the actual loudness values. The x-axis represents the nine estimation frames for each phone.

Table 1: Network estimation results (average absolute error) for pitch and loudness — two male speakers. The pitch values are in Hertz (average percent error) and the loudness values are average phon. The values for [t] are for the release phase only. The term "sonorant" refers to voiced, continuant consonants.

	Pitch (%)		Loudness (phon)	
Speaker	MK	MV	MK	MV
Voiced	1.66	2.07	2.61	3.22
Vowel	1.39	2.01	1.76	2.50
Sonorant	1.76	1.88	3.05	3.45
Voiced Stop	-	-	4.59	3.56
Unvoiced	-	-	3.66	4.45
Fricative	-	-	2.55	3.28
Unvoiced Stop	-	-	3.18	3.39
[*a*]	1.40	2.18	1.37	1.76
[l]	1.18	1.74	2.48	2.30
[s]	-	-	2.33	2.53
[t]	-	-	3.28	2.32

All errors are reported as average absolute error: per cent for pitch and phon for loudness.

Since these networks were designed to predict only local variation in pitch, global variations were removed from the training data. This was accomplished by setting a reference level of 100 Hz for each phone's pitch. Therefore the networks were shown only local variations around 100 Hz and were not subjected to global variations of pitch.

We have run some preliminary tests on sentence material (160 phonetically balanced sentences from two speakers) and the results seem to be promising. In order to produce signals on the sentence scale some other method has to be utilized to produce the larger scale variation. One such possibility is to use, e.g., the Fujisaki algorithm [2]. The results from the micro-level networks are then superimposed on the smoother signals produced by the algorithm for the final estimate. Figure 2 depicts an example where the microprosodic pitch-contour has been predicted by a neural network.

5. CONCLUSION

We have presented some of our ongoing research of Finnish prosody. Our results this far show that the neural network model is applicable to both lexical and microscopic variations of the prosodic parameters. The networks are capable of rule-like behavior and the next, obvious, step is to study the networks themselves to find out more about the factors that govern them and thus the behavior of the parameters they model.

6. REFERENCES

1. R. Aulanko. Microprosodic Features in Speech: Experiments on Finnish. In O. Aaltonen and T. Hulkko, editors, *Fonetiikan Päivät — Turku 1985*, Publications of the Department of Finnish and General Linguistics of the University of Turku, pages 33 – 54, 1985.
2. H. Fujisaki and S. Ohno. Prosodic parameterization of spoken japanese based on a model of the generation process of F0 contours. In *Proc. ICSLP '96*, volume 4, pages 2439–2442, Philadelphia, PA, Oct. 1996.
3. M. Karjalainen and T. Altosaar. Phoneme Duration Rules for Speech Synthesis by Neural Networks. In *Proceedings of the European Conference on Speech Technology*, 1991.
4. M. Karjalainen and T. Altosaar. An object-oriented database for speech processing. In *Proceedings of the European Conference on Speech Technology*, 1993.
5. I. Lehiste and G. Peterson. Vowel Amplitude and Phonemic Stress in American English. *Journal of the Acoustical Society of America*, 31(4):428–435, 1959.
6. M. Vainio and T. Altosaar. Pitch, Loudness, and Segmental Duration Correlates: Towards a Model for the Phonetic Aspects of Finnish Prosody. In H. T. Bunnell and W. Idsardi, editors, *Proceedings of ICSLP 96*, volume 3, pages 2052–2055, Philadelphia, 1996.
7. M. Vainio and T. Altosaar. Pitch, Loudness and Segmental Duration Correlates in Finnish Prosody. In S. Werner, editor, *Nordic Prosody, Proceedings of the VIIth Conference, Joensuu 1996*, pages 247 – 255. Peter Lang, 1998.
8. M. Vainio, T. Altosaar, M. Karjalainen, and R. Aulanko. Modeling Finnish Microprosody for Speech Synthesis. In A. Botinis, G. Kouroupetroglou, and G. Carayannis, editors, *ESCA Workshop on Intonation: Theory, Models and Applications, September 18-20, 1997, Athens, Greece*, pages 309 – 312. ESCA, University of Athens, 1997.
9. E. Vilkman, O. Aaltonen, I. Raimo, P. Arajärvi, and H. Oksanen. Articulatory hyoid-laryngeal changes vs. cricothyroid muscle activity in the control of intrinsic F_0 of vowels. *Journal of Phonetics*, 17:193 – 203, 1989.
10. D. Whalen and A. Levitt. The Universality of Intrinsic F_0 of Vowels. *Journal of Phonetics*, 23:349 – 366, 1995.