https://doi.org/10.3176/lu.1998.3.08

*MATTI KARJALAINEN, TOOMAS ALTOSAAR, MARTTI VAINIO* (Espoo)

# FINNISH SPEECH SYNTHESIS USING WARPED LINEAR PREDICTION AND NEURAL NETWORKS

A text-to-speech synthesis technique, based on warped linear prediction (WLP) and neural networks, is presented for high-quality individual sounding synthetic speech. Warped linear prediction is used as a speech production model with wide audio bandwidth yet with highly compressed control parameter data. An excitation codebook, inverse filtered from a target speaker's voice, is applied to obtain individual tone quality. A set of neural networks, specialised to yield synthesis control parameters from phonemic input in specific contexts, generate the detailed parametric controls of WLP. Neural nets are also used successfully to compute the prosodic parameters. We have applied this approach in prototyping improved text-to-speech synthesis for the Finnish language.

## 1. Introduction and motivation

After a long period of successful developments in text-to-speech (TTS) synthesis, voice quality still remains a challenge. No practical technique yields wide audio bandwidth, near human quality, and individual sounding speech.

Our effort in this study was to find a strategy to improve TTS synthesis for the Finnish language. Earlier achievements were first based on traditional formant synthesis with rule-based control, SYNTE 2 and 3 (Karjalainen, Laine, Toivonen 1980), and then concatenation synthesis called microphonemic synthesis (Lukaszewicz, Karjalainen 1987) similar to the PSOLA technique (Moulines, Carpenter 1990). Concatenative synthesis, based on samples from human speech, easily captures the features from individual speakers. In order to approach full naturalness, however, a huge inventory of samples in different contexts is needed. The algorithms to select concatenative units and to join them in synthesis tend to become complex.

Source-filter models for speech synthesis, such as those used in linear prediction, have more flexibility and allow for easy analysis of control data. The problem remains how to code the excitation (source) and the filter control parameters in a compact way and be able to recompute them from phonemic/phonetic information. Hand-tuned rules and tables, as used in early synthesis, cannot produce the highest quality of speech. Tables of parameter trajectories have similar problems as concatenative synthesis: the size of such inventories grows beyond practical limits when contextual details are included. Among the techniques that are used to

compress and generalise control parameter information through learning are, e.g., neural networks, Hidden Markov Models, and fuzzy or neuro-fuzzy rule systems.

The requirements dictating the choice of methods in our study were to obtain very high quality individual sounding synthesis, wide audio bandwidth (>10 kHz), easy automation of tuning the synthesis to individual speakers using a speech database, moderate memory and processor requirements in implementation, easy integration of audio and visual synthesis (talking head), and preferably as much language independence as possible.

We first discarded the waveform concatenation methods due to the complexity of sample collection and even more due to the difficulty of controlling the detailed contextual effects. An LPC-like source-filter model was found to be more attractive. The success of this approach depends on several factors. A relatively small inventory of source excitations for the synthesis of all phones in the target language should be easily acquirable. The filter parameters should be represented compactly in a form that is suitable to automatic training, e.g., using neural nets.

The problem of ordinary linear prediction with wide bandwidths is that a high filter order is required and the high-frequency portion reserves too much resolution. For example, with a sampling rate of 22 kHz, the traditional rule of thumb leads to an LP filter order of about 24 and most of the filter parameters focus on frequencies above the important formant range below 3.4 kHz (Markel, Gray 1976). This problem was elegantly solved in our case by adopting warped linear prediction (WLP) (Laine, Karjalainen, Altosaar 1994), utilising non-uniform frequency resolution and allowing moderate filter orders of 10—14 almost independently of the sampling rate.

The compactness of synthesis parameter information helped in modelling the generation of these parameters from phonemic input data. Neural networks have been shown to perform this mapping but not without problems. Possible candidates of neural nets are multilayer feedforward nets with phoneme string and synthesis position input, time delay neural networks (TDNN) with time frame input, and recurrent networks, see Karaali et al. (1997) and references in it. Our experience with neural nets has shown that for detailed modelling, specialisation of nets is useful so that each individual net is applied only in a specific context.

In this paper the main features of our approach are described. We have studied the level of voice quality achievable using WLP and specialised neural nets. A full scale synthesiser is under development but already the experiments indicate that a very natural and individual sounding TTS synthesis, practical for implementation, can be obtained.

## 2. Warped linear prediction

The first systematic formulation of warped linear prediction was presented in 1980 (Strube 1980). Later, U. K. Laine, M. Karjalainen and T. Altosaar (1994) have studied various formulations of efficient WLP. The idea of a warped frequency scale and related resolution is based on using allpass sections instead of unit delays in DSP structures. With a proper warping the frequency scale shows a good match to the psychoacoustically defined Bark scale (Smith, Abel 1995) thus optimising the frequency resolution from the point of view of auditory perception. The filter structure shown in Figure 1 has been used in our WLP synthesis experiments.

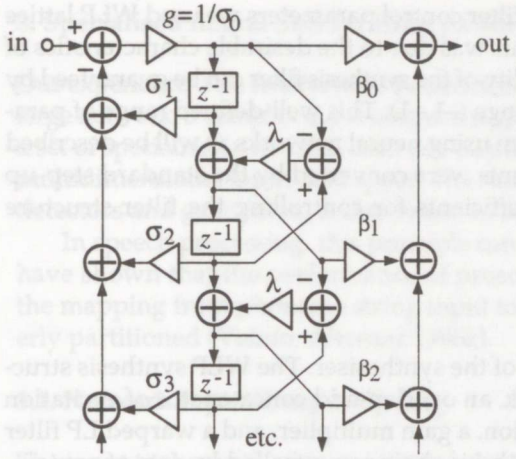


*Figure 1.* **A realisable WIIR structure with first-order allpass delays and a single unit delay.**

The advantage gained when using Bark warping is that in wideband synthesis the filter order can be reduced remarkably without sacrificing the frequency resolution at low frequencies. At high frequencies the spectral resolution is worse, nevertheless this is exactly how hearing functions. We have experimentally evaluated the voice quality of WLP and normal LP for various filter orders when the sampling rate is 22 kHz. Ordinary LP yields good quality with orders of 20—24 while WLP works comparably with orders of 10—14. Figure 2 shows synthesis filter responses for a vowel spectrum (Finnish /a/) using ordinary LP and WLP.

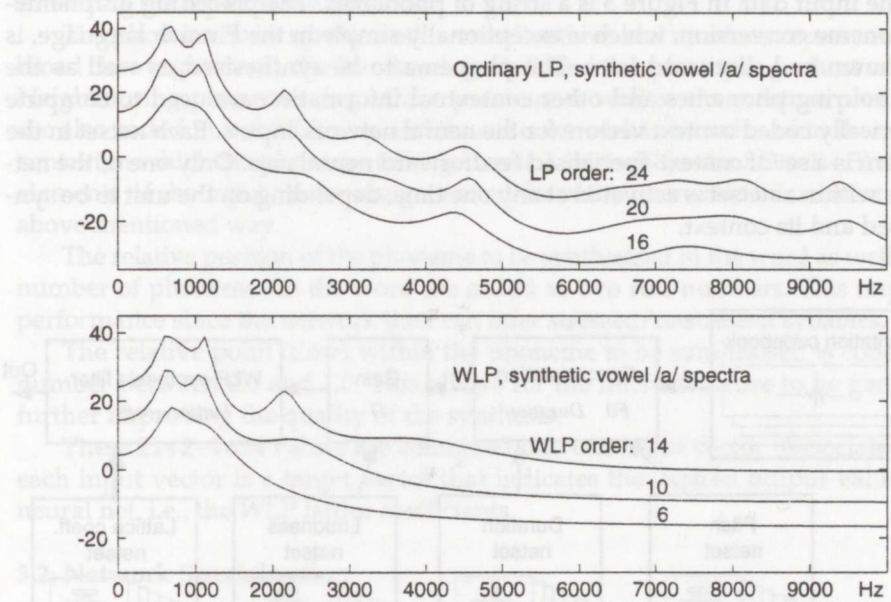


*Figure 2.* **LP and WLP spectra of vowel /a/ for different filter orders.**

The main advantage of WLP over LP is the compression of control parameter data which helps in the training of neural nets to generate these parameters. A lower filter order is also advantageous for fast computation but this is counteracted by the inherently more complex structure of the warped IIR filters (Figure 1). It is also possible to expand the WIIR filter structure into an ordinary direct form IIR filter but the WIIR structure is numerically more robust as discussed in Karjalainen, Härmä, Laine 1996. Since on modern processors (DSPs, Pentium, PowerPC) such filters consume only a few per cent of CPU resources, the robust and straightforward WIIR structure of Figure 1 has been used in our synthesiser.

As a final representation for WLP filter control parameters we used WLP lattice coefficients (reflection coefficients). This was due to the desirable characteristics of reflection coefficients whereby the stability of the synthesis filter can be guaranteed by limiting the coefficient values in the range (–1,+1). This well-defined range of parameters also helps when generating them using neural networks as will be described below. The warped reflection coefficients were converted by the standard step-up procedure to warped polynomial coefficients for controlling the filter structure shown in Figure 1.

### 2.1. System Configuration

Figure 3 illustrates the block diagram of the synthesiser. The WLP synthesis structure consists of an excitation codebook, an overlap-add concatenator of excitation signals for pitch and duration generation, a gain multiplier, and a warped LP filter (WIIR synthesis filter). This voice synthesis chain is controlled by sets of context-specialised neural networks (netsets), for filter parameters, pitch, duration, and gain controls. Neural network inputs as well as the selection of a proper network within a netset is based on the phoneme to be synthesised, its phonemic context as well as other contextual information.

The input data in Figure 3 is a string of phonemes. The preceding grapheme-to-phoneme conversion, which is exceptionally simple in the Finnish language, is not shown and discussed here. The phoneme to be synthesised as well as the neighbouring phonemes and other contextual information are used to compute numerically coded context vectors for the neural network inputs. Each netset in the diagram is a set of context-specialised feedforward neural nets. Only one of the networks within a netset is activated at any one time, depending on the unit to be synthesised and its context.

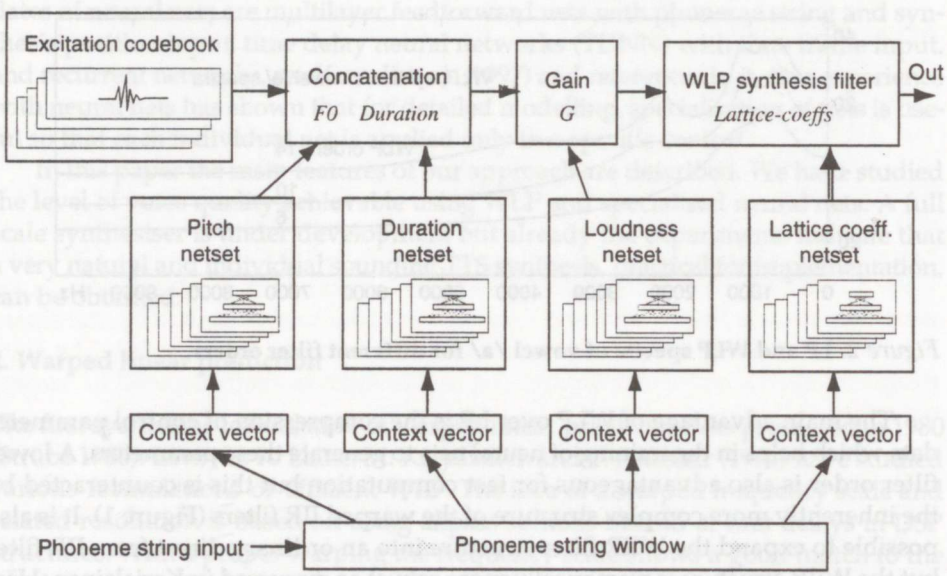


*Figure 3.* **Configuration of the speech synthesis system using warped linear prediction and specialised neural nets.**

## 3. Specialised neural nets for filter parameter control

Our experience with feedforward neural nets has shown that, instead of using a single large network, a complex input-output mapping is more easily and precisely learnt by a set of specialised networks, each one contributing only within a specific region of a multidimensional input data space. The same strategy, the utilisation of specialised detectors and generators, is also found in biological and human neural systems.

In speech processing, this principle can be utilised in various ways. Earlier we have shown that the performance of prosodic feature models is improved when the mapping from phoneme string input to duration, pitch, or signal gain is properly partitioned (Vainio, Altosaar 1996a).

### 3.1. Network Input/Output Coding

The input to the synthesiser consists of phonemic information (a string of phonemes converted from a string of graphemes) as well as phonetic information (e.g., factors affecting prosody indicated by punctuation). This symbolic information must be converted into numerical form to allow neural networks to be utilised in the generation of synthesis control parameters. We have used three types of information to constitute the input to the networks:

The phoneme to be synthesised is coded as three real numbers representing the broad class (e.g., vowel), the fine class (e.g., /a/), and the quantity (e.g., short vs. long). Neighbouring phonemes (e.g., three previous as well as three future phonemes) are also coded in a similar way and thus the network is introduced to the specific context in which the phoneme to be generated exists. Therefore $(3 + 1 + 3) \times 3 = 21$ elements of the input vector are generated from the phonemic information in the above mentioned way.

The relative position of the phoneme to be synthesised in the word as well as the number of phonemes in the word are coded as two real numbers. This improves performance since the network then can infer stressed/unstressed syllables.

The relative point (time) within the phoneme to be synthesised is coded as a number between 0.0 and 1.0. This allows for the microstructure to be generated further improving the quality of the synthesis.

These 21+2+1=24 values are combined into one input vector. Associated with each input vector is a target vector that indicates the desired output values of a neural net, i.e., the WLP lattice coefficients.

### 3.2. Network Specialisation

Phoneme networks model the WLP coefficients at any temporal point within a phoneme. However, when moving across phoneme boundaries, switching in a new network may cause discontinuities to occur in the coefficients. To achieve more smooth transitional performance around these areas a set of diphone WLP synthesis networks are taught and utilised in a manner similar to the phoneme nets. Amplitude mixing (cross-fading) the outputs of both network types improves the quality of synthesis.

Table 1 shows the average absolute error of the lattice coefficients for a set of WLP diphone synthesis networks as a function of the degree of specialisation. As specialisation decreases the error increases. As an example of spectral error due to lattice coefficient error, Figure 4 displays the WLP spectrum slightly past the diphthong transition [e]—[i] in the Finnish word /keinu/. The topmost curve represents the

actual WLP spectrum at this point in the signal while the other curves (in order of decreasing specialisation) represent the synthesised spectra using the networks listed in Table 1. The [e]—[i] specific network produces the most accurate spectral estimate (second topmost curve).

*Table 1*

**Lattice coefficient error vs. network specialisation**

| Specialisation | Diphone Type | Coefficient Error |
|---|---|---|
| specific | /e/—/i/ | 5.0% |
| ... | front vowel — front vowel | 5.3% |
| ... | vowel — front | 6.1% |
| general | any—any | 7.5% |

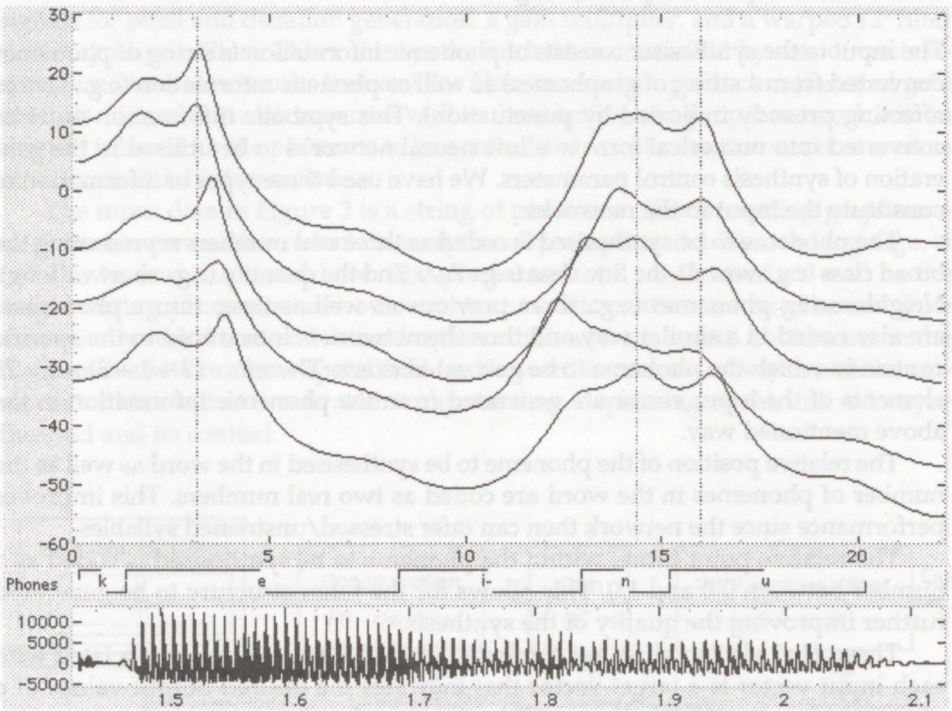


*Figure 4.* **WLP spectra (dB vs. Bark scale) at a certain time instant in an [e]—[i] transition of word /keinu/. The top curve is the target spectrum and the other ones are neural net generated cases (Table 1) in order of decreasing specialisation.**

### 3.3. Speech Database and Network Training

The speech material used for training and evaluating the networks consisted of approximately 2000 Finnish words spoken in isolation by a single male speaker. This manually segmented and phonetically transcribed material was divided into training and evaluation sets with a 2:1 ratio on a word basis. Each phone or diphone segment in either the training or evaluation set provided for 13 temporally nonlinearly spaced training elements. The number of elements in the training and evaluation sets for the most general diphone network exceeded 100,000 and 50,000, respectively. As the degree of specialisation increased the size of the sets decreased.

For each degree of specialisation the number of hidden nodes was systematically varied to determine the optimum network size so as to match the network to the difficulty of the mapping problem. Three hidden nodes was found to minimise the error for the most specialised network while the more general networks performed better with a substantially larger number of nodes. For example, the any—any diphone net displayed in Figure 4 utilised 500 hidden nodes and this explains the relatively high level of spectral detail produced by this network.

## 4. Excitation codebook

The excitation codebook is an indexed table of residual signals, extracted from the speech database signal entries for the individual speaker to be modelled. In the most simple case a single excitation pattern may be used for all voiced sounds. However, a more natural voice quality is obtained if each phoneme has a different entry in the codebook, each representing a typical case of this specific phoneme. If desired, the codebook can be made even more specialised, e.g., by providing a separate entry for some critical allophones.

The entries of the excitations are concatenated during synthesis so that the desired pitch is generated according to the pitch target produced by the corresponding net-set. For unvoiced sounds, white noise is used as an excitation signal.

## 5. Prosody control

Prosody control is accomplished with three sets of networks for segmental durations, fundamental frequency, and gain (loudness). Their input is similar to the WLP networks' input with some difference in the phonetic information. Pitch nets are coded onto the semitone scale, loudness nets onto the phon scale, and the duration nets onto a logarithmic time scale. Again, specialisation is utilised.

Our prosody control results were as follows: duration estimation was the most difficult task and specialisation was needed for the error to decrease below 20%, the difference limen. A 2.2 phon error was achieved with loudness networks — one phon is generally considered just noticeable. An error of 3.5% was measured for the pitch networks: this amounts to about 0.6 semitones at 100 Hz and is well below the 1.5 to 2 semitone threshold for speech ('t Hart, Collier, Cohen 1990). Prosody control is discussed in more detail in Karjalainen, Altosaar 1991; Vainio, Altosaar 1996a; 1998.

## 6. Summary

An experimental framework for individual sounding TTS utilising WLP and specialised neural network sets for controlling spectral and prosodic parameters has been presented. The system described in this paper is in the development stage and so far has been trained and evaluated on isolated words. Future work includes extending the synthesiser to the sentence level as well as implementing a real-time version.

## REFERENCES

K a r a a l i, O. et al. 1997. Text-to-Speech Conversion with Neural Networks: A Recurrent TDNN Approach. — Proceedings of EUROSPEECH'97, Rhodes.

K a r j a l a i n e n, M., L a i n e, U. K., T o i v o n e n, R. 1980. Aids for the Handicapped Based on SYNTE 2 Speech Synthesiser. — Proceedings of IEEE ICASSP-80, Denver.

Karjalainen, M., Altosaar, T. 1991. Phoneme Duration Rules for Speech Synthesis by Neural Networks. — Proceedings of the European Conference on Speech Technology EUROSPEECH'91, Genoa.

Karjalainen, M., Härmä, A., Laine, U. K. 1996. Realizable Warped IIR Filter and Their Properties. — Proceedings of IEEE ICASSP-96, Munich.

Laine, U. K., Karjalainen M., Altosaar T. 1994. Warped Linear Prediction (WLP) in Speech and Audio Processing. — Proceedings of IEEE ICASSP-94, Adelaide.

Lukaszewicz, K., Karjalainen, M. 1987. Microphonemic Method of Speech Synthesis. — Proceedings of IEEE ICASSP-87, Dallas.

Markel, J. D., Gray, A. H. 1976. Linear Prediction of Speech. New York.

Moulines, E., Carpenter, F. 1990. Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. — Speech Communication, 9 (5/6), 453—467.

Smith, J. O., Abel, J. S. 1995. The Bark Bilinear Transform. — Proceedings of IEEE ASSP Workshop, Mohonk—New Paltz.

Strube, H. W. 1980. Linear Prediction on a Warped Frequency Scale. — Journal of the Acoustical Society of America, 68, 1071—1076.

't Hart, J., Collier, R., Cohen, A. 1990. A Perceptual Study of Intonation. Cambridge.

Vainio, M., Altosaar, T., 1996a. Pitch, Loudness, and Segmental Duration Correlates. Towards a Model for the Phonetic Aspects of Finnish Prosody. — Proceedings of ICSLP'96, Philadelphia.

—— 1998. Pitch, Loudness, and Segmental Duration Correlates in Finnish Prosody. — Nordic Prosody. Proceedings of VIIth Conference. Joensuu, 1996, Frankfurt am Main.