Electronics and Communications in Japan, Part 3, Vol. 80, No. 9, 1997
Translated from Denshi Joho Tsushin Gakkai Ronbunshi, Vol. J79-A, No. 7, July 1996, pp. 1227–1235

Pitch Extraction and Voiced/Unvoiced Detection of Speech by Cross-Coupling Multi-Layered Neural Network with Feedback Architecture

Hideo Miyabayashi

Toyama National College of Maritime Technology, Shinminato, Japan 933-02

Tetsuo Funada

Faculty of Engineering, Kanazawa University, Kanazawa, Japan 920

SUMMARY

Pitch frequency is one of the most important voice characteristics, and its accurate extraction is important not only in speech analysis and synthesis, but also in speech coding, speech recognition, speaker recognition, and the like. Existing methods of improving extraction accuracy include waveform processing, correlative processing, and spectral processing. This paper describes the use of a neural network to extract pitch from voice features delivered from the bandpass filter pairs (BPFPs) proposed by Fonda et al. Three types of multi-layered neutral networks able to learn time-continuity and high accuracy discrimination functions and have a recurrent structure are tested. The cross-coupling multi-layered neural network with feedback architecture gives the best improvement over conventional neural networks, and exhibits superior ability for learning time continuity of pitch and U/V information. © 1997 Scripta Technica, Inc. Electron Comm Jpn Pt 3, **80**(9): 48–58, 1997

Key words: Speech detection; pitch extraction; multilayer neural network; feedback architecture; cross-coupling neural network.

1. Introduction

Pitch frequency is one of the most important basic parameters of the voice. Accurate extraction of the voice pitch is not limited to speech analysis and synthesis systems, but it is also a fundamental and important subject matter for speech coding, speech recognition, speaker recognition, and so on. Typical pitch extraction methods that hitherto have been investigated to improve the accuracy include those based on wave form processing, correlative processing, and spectral processing, but a decisive method has not yet been established [12].

Recently, several pitch extraction methods using neural networks (NNs) have been reported [2–6]. In this paper, using an NN proposed here, we extract the pitch from the voice features delivered through the use of bandpass filter pairs (BPFPs) proposed by Funada et al. [7].

We teach pitch extraction and unvoiced/voiced (U/V) detection functions to three types of the multi-layered NNs that are able to learn time-continuity and high-accuracy discrimination functions, and that have a recurrent structure with feedback loops from the output layer and cross-coupling paths within each hidden layer, and we compare their detectability. The experimental results showed clearly that the cross-coupling multi-layered NN with feedback archi-

CCC1042-0967/97/090048-11

tecture (hereafter called CCNN-F) has the highest improved accuracy when compared with conventional NNs, and that learning abilities of the time continuity of pitch and U/V information, as well as discrimination functions were much better obtained by cross-coupling paths within each hidden layer and feedback loops from the output layer to the first hidden layer.

Now, the mechanism by which the frequency resolution becomes sharper, with passing through the repeater of the signal in the centripetal system ascending from the auditory nerve to the corpus geniculatum medial of the diencephalon has been clarified [12]. At the network level, this mechanism suggests the existence of excitatory and inhibitory cross-couplings among the units within a same layer. Furthermore, we see that the centrifugal system descending from the cerebral cortex is also found [12], thereby suggesting the existence of feedback couplings in order to adapt to the input voice.

In what follows, we first explain the NN used in this research and its learning algorithm and then the pitch extraction system. Thereafter, we discuss in detail the comparative performance results.

2. Coupling Topology of Neural Networks and Learning

We explain here four types of the four-layered NN used in the present research and these learning algorithms.

2.1. Neural network topologies

(1) Simple multi-layered NN (net topology: 000)

Figure 1(a) shows this NN structure. This is the conventional feed-forward structure.

(2) Multi-layered NN with cross-coupled hidden layers (net topology: 010)

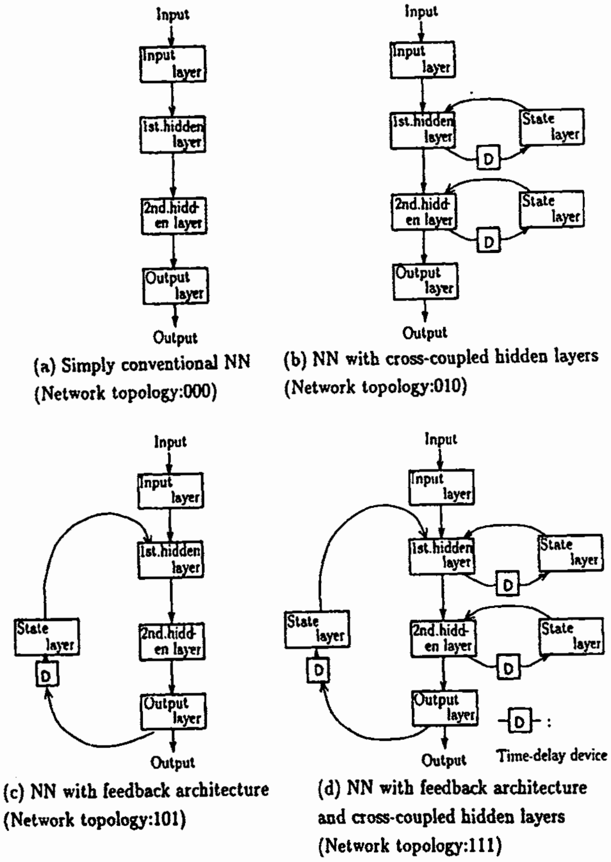
The structure of this NN is shown in Fig. 1(b). This one organizes cross-coupled hidden layers NN by inserting into the hidden layers of the conventional multi-layered NN a state layer in order to preserve the middle-layer state of a previous time step. This state is then propagated to each unit within the same layer through a time-delay element and the state layer, thereby mutually cross-coupling all the units within each middle layer (including a unit with itself). In other words, it is the cross-coupling multi-layered NN structure.

(3) Feedback coupling multi-layered NN without cross-coupled hidden layers (net topology: 101)

The structure of this NN is shown in Fig. 1(c). Feedback coupling is done between the output layer and a hidden layer (the first one) by adding a state layer to the conventional multi-layered NN in order to preserve the output layer state of a previous time step, and propagating this state to each unit of the first hidden layer through a time-delay element and the state layer.

(4) Feedback coupling multi-layered NN with cross-coupled hidden layers (net topology: 111)

The structure of this NN is shown in Fig. 1(d). This NN is a combination of the two structures of Figs. 1(b) and 1(c). This model is designated as CCNN-F. It is possible to realize this one by an NN with feedback and cross-coupling with only time-delay elements and absolutely no state layers. However, ours is an extended representation in which the past information of the state layer is referable in general.



(a) Simply conventional NN (Network topology:000)

(b) NN with cross-coupled hidden layers (Network topology:010)

(c) NN with feedback architecture (Network topology:101)

(d) NN with feedback architecture and cross-coupled hidden layers (Network topology:111)

Fig. 1. Conventional NN and proposed NNs with feedback architecture or cross-coupled hidden layers.

2.2. Learning algorithm

The learning algorithm used in this research follows the back-propagation training method [8]. Learning of the feedback coupling weights can be realized with the conventional error back-propagation technique by considering the observed pattern of the output layer in the previous time step through the state layer as a pseudo-external input pattern. Learning of the cross-coupling weights takes place by an approximate error back-propagation technique, where the learning error of the hidden layer in the previous time step is propagated only from within the same layer directly through the state layer. We explain below the learning algorithm of the CCNN-F model as an example.

Let I, $H1$, $H2$, and O be the number of units in the input, first hidden, second hidden and output layers of the present model, respectively. The input-output relation of a unit in the NN may be formally expressed in general as

$$x_j^{(n+1)}(t) = \sum_i w_{ji}^{(n)} y_i^{(n)}(t) + \theta_j^{(n+1)} \quad (1)$$

$$y_j^{(n+1)}(t) = f(x_j^{(n+1)}(t)) \quad (2)$$

Here, t represents the time progression by discrete integer steps. In general, $f(x_j^{(n+1)}(t))$ is given by the following sigmoid function,

$$f(x_j^{(n+1)}(t)) = \frac{1}{1 + \exp(-x_j^{(n+1)}(t)/\kappa)} \quad (3)$$

where κ is the sigmoid slope coefficient.

In the first hidden layer, which has feedback couplings from the output layer and self cross-couplings, Eqs.(1) and (2) become

$$x_j^{(2)}(t) = \sum_{i=1}^{I} w_{ji}^{(1)} y_i^{(1)}(t) + \sum_{m=1}^{O} u_{jm} y_m^{(N)}(t-1) + \sum_{l=1}^{H1} v_{jl}^{(1)} y_l^{(2)}(t-1) + \theta_j^{(2)} \quad (4)$$

$$y_j^{(2)}(t) = f(x_j^{(2)}(t)) \quad (5)$$

Here, $y_m^{(N)}(t-1)$ is the output value of the m-th output unit in the previous time step, u_{jm} is the value of a feedback coupling weight from the m-th output unit in the previous time step to the j-th unit of the first hidden layer, $y_l^{(2)}(t-1)$ is the previous value of the l-th unit in the first hidden layer, and $v_{jl}^{(1)}$ is the value of a cross-coupling weight from the l-th unit of the first hidden layer to the j-th unit of the same layer. Further, if the second term of Eq. (4) is considered as a pseudo-external input, it can be treated like the first term of the same expression.

In the second hidden layer there are only cross couplings within the same layer, so Eqs. (1) and (2) become

$$x_j^{(3)}(t) = \sum_{i=1}^{H1} w_{ji}^{(2)} y_i^{(2)}(t) + \sum_{l=1}^{H2} v_{jl}^{(2)} y_l^{(3)}(t-1) + \theta_j^{(3)} \quad (6)$$

$$y_j^{(3)}(t) = f(x_j^{(3)}(t)) \quad (7)$$

In the output layer, Eqs. (1) and (2) become

$$x_j^{(N)}(t) = \sum_{i=1}^{H2} w_{ji}^{(3)} y_i^{(3)}(t) + \theta_j^{(N)} \quad (8)$$

$$y_j^{(N)}(t) = f(x_j^{(N)}(t)) \quad (9)$$

As for the back-propagation training, weights are adjusted to make the sum **E** of the squared output errors in Eq. (10) become a minimum.

$$\mathbf{E} = \sum_t E(t) = \frac{1}{2} \sum_t \sum_{j=1}^{O} (y_j^{(N)}(t) - y_j^d(t))^2 \quad (10)$$

Here, $E(t)$ is the sum of the squared output errors at each time step t in the learning cycle and $y_j^d(t)$ is the target output value of the j-th output unit.

Weights are adjusted in each time interval (sequential adjusting law) and the adjustments of weight and bias are directly proportional to the partial derivatives of $E(t)$, as indicated in the following equations

$$\left.\begin{aligned} \Delta w_{ji}^{(n)}(t) &= -\eta_1 \frac{\partial E(t)}{\partial w_{ji}^{(n)}} + \alpha \Delta w_{ji}^{(n)}(t-1) \\ \Delta u_{jm}(t) &= -\eta_1 \frac{\partial E(t)}{\partial u_{jm}} + \alpha \Delta u_{jm}(t-1) \\ \Delta v_{jl}^{(n)}(t) &= -\eta_1 \frac{\partial E(t)}{\partial v_{jl}^{(n)}} + \alpha \Delta v_{jl}^{(n)}(t-1) \end{aligned}\right\} \quad (11)$$

$$\Delta \theta_j^{(n+1)}(t) = -\eta_2 \frac{\partial E(t)}{\partial \theta_j^{(n+1)}} + \alpha \Delta \theta_j^{(n+1)}(t-1) \quad (12)$$

where, $\Delta \omega_{ji}$, Δu_{jm}, and Δv_{jl} are the adjusting quantities for the respective weights; $\Delta \Theta_{ij}$ is the adjusting quantity of the bias; η_1 and η_2 are the learning rates used for the weight and bias; and α is the momentum.

The partial derivatives used in the first terms of Eqs. (11) and (12) can be expressed, respectively, by

$$\left.\begin{aligned} \frac{\partial E(t)}{\partial w_{ji}^{(n)}} &= -\delta_j^{(n+1)}(t)\frac{\partial x_j^{(n+1)}(t)}{\partial w_{ji}^{(n)}} \\ \frac{\partial E(t)}{\partial u_{jm}} &= -\delta_j^{(n+1)}(t)\frac{\partial x_j^{(n+1)}(t)}{\partial u_{jm}} \\ \frac{\partial E(t)}{\partial v_{jl}^{(n)}} &= -\delta_j^{(n+1)}(t)\frac{\partial x_j^{(n+1)}(t)}{\partial v_{jl}^{(n)}} \end{aligned}\right\} \quad (13)$$

$$\frac{\partial E(t)}{\partial \theta_j^{(n+1)}} = -\delta_j^{(n+1)}(t)\frac{\partial x_j^{(n+1)}(t)}{\partial \theta_j^{(n+1)}} \quad (14)$$

Here, $-\delta_j^{(n+1)}(t)$ is

$$-\delta_j^{(n+1)}(t) = \frac{\partial E(t)}{\partial y_j^{(n+1)}(t)} f'(x_j^{(n+1)}(t)) \quad (15)$$

where the derivative of the sigmoid function $f(x_j^{(n+1)}(t))$ is

$$\begin{aligned} f'(x_j^{(n+1)}(t)) &= f(x_j^{(n+1)}(t)) \\ &\quad \times(1 - f(x_j^{(n+1)}(t)))/\kappa \end{aligned} \quad (16)$$

Let us consider for simplicity the error back-propagation from other units only by direct coupling paths (approximate learning technique [9, 10]. From Eqs. (4), (6), and (8), the partial derivatives used in the right side of Eqs. (13) and (14) become

$$\left.\begin{aligned} \frac{\partial x_j^{(n+1)}(t)}{\partial w_{ji}^{(n)}} &= y_i^{(n)}(t) \\ \frac{\partial x_j^{(n+1)}(t)}{\partial u_{jm}} &= y_m^{(N)}(t-1) \\ \frac{\partial x_j^{(n+1)}(t)}{\partial v_{jl}^{(n)}} &= y_l^{(n+1)}(t-1) \end{aligned}\right\} \quad (17)$$

$$\frac{\partial x_j^{(n+1)}(t)}{\partial \theta_j^{(n+1)}} = 1 \quad (18)$$

From Eq. (15), the error $\delta_j^{(N)}(t)$ of the j-th output unit may be expressed by

$$\delta_j^{(N)}(t) = f'(x_j^{(N)}(t))(y_j^d(t) - y_j^{(N)}(t)) \quad (19)$$

The error $\delta_j^{(3)}(t)$ of the j-th unit of the second hidden layer by

$$\begin{aligned} \delta_j^{(3)}(t) = f'\ (x_j^{(3)}(t))(&\sum_{k=1}^{O} \delta_k^{(N)}(t) w_{kj}^{(3)} \\ &+ \sum_{l=1}^{H2} \delta_l^{(3)}(t+1) v_{lj}^{(2)}) \end{aligned} \quad (20)$$

and the error $\delta_j^{(2)}(t)$ of the j-th unit of the first hidden layer by

$$\begin{aligned} \delta_j^{(2)}(t) = f'\ (x_j^{(2)}(t))(&\sum_{k=1}^{H2} \delta_k^{(3)}(t) w_{kj}^{(2)} \\ &+ \sum_{l=1}^{H1} \delta_l^{(2)}(t+1) v_{lj}^{(1)}) \end{aligned} \quad (21)$$

All the adjustments of weight and bias can be determined by the learning rules with the above Eqs. (11) through (18).

3. Pitch Extraction System

The pitch extraction system used in the present research is composed of a BPFP bank block, a pitch extraction NN block, and a U/V detection NN block. We explain each block in turn.

3.1. BPFP bank block

The BPFP bank extracts the voice features of a frame unit (period 10 ms, length 30 ms) from a sampled voice signal (0.1 ms interval, 16 bit resolution) by means of the BPFP method. The BPFP method determines the slope and power level for the frequency power spectrum at various frequency points of the objective frequency band pass (in this research, 11 channel banks at 15 Hz intervals, with center frequency 100 Hz to 250 Hz, 11 channel banks at 30 Hz intervals, with center frequency 280 Hz to 580 Hz). The result is the features vector of the voice. What we call slope is a value expressing the degree and direction of distance of the higher harmonics close to the center frequency of each channel from such center frequency. When the higher harmonic frequency is larger than the center frequency it has a positive value, and a negative one when it is smaller [7].

The BPFP bank delivers such 44-dimensional features vectors for each frame.

3.2. Normalization of the features vector

In order to carry out the pitch extraction and U/V detection independently of the voice signal amplitude, it is necessary to normalize the features vector delivered by the BPFP bank block in the preceding processing part.

The 22-dimensional features vector based on the power level for the frequency spectrum is normalized in the range [0, 1] by determining for each frame the largest valued component and dividing all the components by this largest value. Also, for the 22-dimensional features vector, according to the slope independently of each amplitude, normalization is done for each frame in the range [−1, 1].

Hence, the inputs to the following pitch extraction NN and U/V detection NN blocks are normalized 44-dimensional feature vectors for each frame of the voice signal.

3.3. U/V detection NN block

The U/V detection NN block detects the U/V of a frame based on the features vector delivered by the BPFP bank. The NN is built with an input layer of 44 units, two middle layers, the first one with 30 units and the second one with 15 units, and an output layer with 1 unit.

When training the NN, a teacher signal is given with 0.99 for voiced frames and 0.01 for unvoiced or silent frames. When evaluating the trained NN registers a voiced frame if the value of the putout layer exceeds 0.5 and an unvoiced one if it its less than 0.5.

3.4. Pitch extraction NN block

For the voiced frame detected in the U/V detection NN block, the pitch extraction NN block extracts the pitch frequency of a frame based on the features vector delivered by the BPFP bank. This NN has the structure as the U/V detection NN.

When training the NN, a teacher signal is given with a correct pitch for the voiced frame. During evaluation, successful pitch extraction is considered to be achieved when the extracted frequency falls within ±5% of the correct pitch.

Considering that the upper bound of the pitch frequencies for the voiced signal is 450 Hz and the lower bound is 50 Hz, the teacher signal for a correct pitch given in the output layer was actually transformed by the formula

$$(teacher's\ signal) = \frac{log((correct\ pitch)/50.0)}{log(9.0)} \quad (22)$$

and thereby normalization is done in the range [0, 1]. We can determine the extracted pitch from the output value of the NN output layer by means of the inverse formula

$$(extracted\ pitch) = 50.0 \times exp((output\ value) \times log(9.0)) \quad (23)$$

4. Experimental Method

Next we explain the makeup of the experiment that evaluated the comparative performance and training of the four types of pitch extraction NNs and U/V detection NNs.

4.1. Speech data

The data used in the experiment were selected from the continuous speech corpus for research (the Acoustic Society of Japan) as follows:

- Trained speakers: six persons, three males (m01, m02, t01) and three females (m11, m12, t11)
- Untrained speakers: two persons, one male (t03) and one female (t12)
- Training data 1: five sentences (a01 through a05) pronounced by each trained speaker for a total of 30 sentences
- Training data 2: five sentences (a06 through a10) pronounced by each trained speaker for a total of 30 sentences
- Non-training data used for evaluation: When learning with training data 1, each trained speaker pronounced sentences a06 through a10 for a total of 30 sentences; in the case of training with training data 2, each trained speaker pronounced sentences a01 through a05 for a total of 30 sentences; the 10 sentences (a01 through a10) were pronounced by each of the non-trained speakers for a total of 20 sentences.

4.2. Correct pitch

The speech data used for training the network was processed through the BPFP bank block, and the correct U/V and pitch information was extracted automatically according to matches with standard patterns from 128 classes (70–451 Hz range, 3 Hz resolution). In addition, the correct pitch was modified visually. Furthermore, in the case of an unvoiced frame, the correct pitch value was set for convenience as 0. The U/V information given as the teacher signal was dependent on the correct pitch value (=0/≠0).

4.3. Pitch extraction experiment

We carried out pitch extraction using each NN trained with the training data sets 1 and 2, separately and jointly, and did a comparative study on the different extraction ability due to differences in network topologies. The pitch extraction NN was evaluated for voiced frames with non-zero correct pitch. The same learning conditions were set up for the training of all four types of the NNs (slope coefficient of the sigmoid function: 0.8; learning rate: initial value 0.8, decreasing rate 0.99; momentum: initial value 0.5, decreasing rate 0.99; weight initial values: –0.5 to +0.5; bias initial values: –0.3 to +0.3; learning cycles: 2000). The evaluation items for the performance of the trained NNs were the following:

- The ratio of frames for which the extracted pitch was greatly different (> ±20%) from the correct pitch (gross pitch error, GPE)
- The ratio of frames with successful pitch extraction (correct rate)

4.4. U/V detection experiment

We carried out U/V detection using each NN trained with the training data sets 1 and 2, separately and jointly, and did a comparative study on the differing detection ability due to differences in network topologies. The same learning conditions were set up for the training of all four types of the NNs (Slope coefficient of the sigmoid function: 1.5, learning cycles: 1000, and all the other parameters as in the pitch extraction experiment). The evaluation items for the performance of the trained NNs were

- The ratio of unvoiced frames mistaken as voiced frames (unvoiced-to-voiced error, UVE),
- The ratio of voiced frames mistaken as unvoiced frames (voiced-to-unvoiced error, VUE),
- The ratio of correct frames (correct rate)

5. Experimental Results and Discussion

We first analyze the learning effects of time continuity and discrimination function by feedback couplings and cross-couplings, from the pitch extraction and U/V detection results for the cases of training with training data sets 1 and 2 separately, and explain next from the pitch extraction and U/V detection results when using both the training data sets jointly.

5.1. Accuracy of the discrimination function

The scattering results of the extraction accuracy of pitch information in the pitch extraction NNs trained with training data sets 1 and 2 separately are shown in Table 1. We use the percentage of GPE, and for non-GPE frames the average and the variance values of the absolute deviations from the correct values. In Table 1 we also have the values evaluated with non-training data by using the NNs trained with training data sets 1 and 2 separately (denoted hereafter as "training data 1 NN" and "training data 2 NN"). It can be seen from this table that in the net topologies 010 and 111 the GPE and the scattering of the extraction accuracy are the lowest for both the training and the non-training data. Figure 2 shows an example of pitch extraction results. We can confirm visually the scattering of the extraction accuracy and the state of errors for each net structure. For the silent or unvoiced sections, the values are not shown here.

The comparison of the pitch extraction and U/V detection correct rate between the net topologies 010, 111 with cross-coupled hidden layers and the net topologies 000, 101 is shown in Table 2. In both the "training data 1 NN" and the "training data 2 NN," the efficacy of the cross-couplings can be notably appreciated in the case of pitch extraction for the well-known and the unknown speakers.

Table 1. Scatter of the accuracy of pitch extraction for each NN
(non-training data, a unit: average [Hz], variance [Hz^2], GPE [%])
Average/variance: average/variance values of the errors in extracted pitches, excluding the GPE.

	Training data 1NN						Training data 2NN					
Net	Well-known speakers			Unknown speakers			Well-known speakers			Unknown speakers		
Topology	Average	Variance	GPE	Average	Variance	GPE	Average	Variance	GPE	Average	Variance	GPE
000	3.03	16.43	1.48	2.73	12.11	0.63	3.25	17.98	1.47	3.09	12.70	0.54
010	2.85	13.73	1.09	2.62	10.04	0.28	2.94	12.84	0.65	2.74	9.32	0.3
101	2.92	18.09	1.61	2.48	10.46	0.65	3.39	18.19	1.64	3.22	13.67	0.54
111	2.87	14.01	1.28	2.68	11.11	0.32	2.95	15.46	0.51	2.80	10.74	0.20

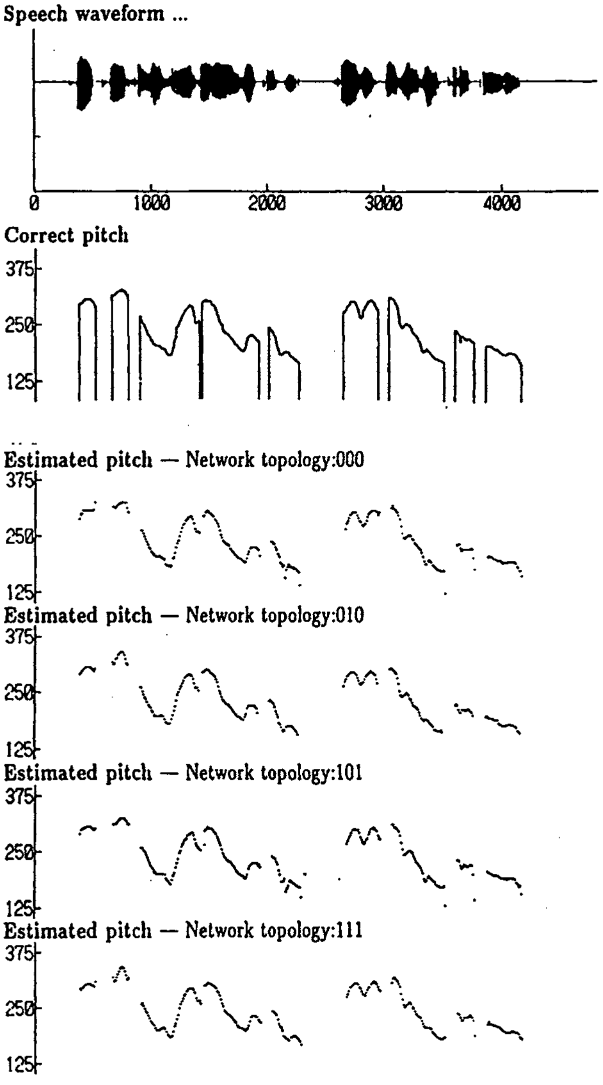


Fig. 2. An example of pitch extraction for female speaker by using each NN trained with data 1(t12-a05, non-training speaker).

From the above facts, it can be said that training of the NNs with cross-coupled hidden layers effectively improves the accuracy of the discrimination function in pitch extraction. This is because the cross-couplings produce a generalization of learning from a small number of training data, by smoothing of the pitch variation information [13].

5.2. Learning effects of time continuity

The learning effects of time continuity of U/V information in the U/V detection NNs trained with the training data sets 1 and 2 separately are shown in Table 3. For the average and variance values in this table, we determined the average/variance of the changes in the estimated value of digitized U/V and the correct one from the respective values in the preceding frame, and divided the average/variance of the estimated value by that of the correct value (increasing rate)—these are the values shown. We see that the increasing rate of the average/variance values are small in the net topologies 101 and 111 for both the well-known and the unknown speakers, i.e., the changing count of the U/V estimated value was close to the count of the correct value. An example of U/V detection results is shown in Fig. 3. We can confirm visually the state of the errors in detection for each net structure.

The comparison of the pitch extraction and U/V detection correct rate between the net topologies 101, 111 with feedback architecture and the net topologies 000, 010 is shown in Table 4. The efficacy of the feedback couplings is different for training data 1 NN and training data 2 NN, but for the unknown speakers, if we reconsider on the average with training data 1 NN and training data 2 NN, we may say that there is a slightly improved effect by the feedback couplings in both the pitch extraction and U/V detection.

Table 2. Comparison of correct rate of pitch extraction and U/V detection by using NN with cross-coupled hidden layers (non-training data, a unit: %). The correct rate values of neural net topologies 010 and 111, relative values to these from neural net topologies 000 and 101 respectively_

		Training data 1NN		Training data 2NN	
	Net topology	Well-known speakers	Unknown speakers	Well-known speakers	Unknown speakers
Pitch extraction	010	+1.95	+1.86	+4.44	+3.84
	111	+0.98	+0.68	+4.77	+4.86
U/V detection	010	+0.76	−0.03	+0.01	+0.14
	111	−0.06	−0.21	+0.19	+0.65

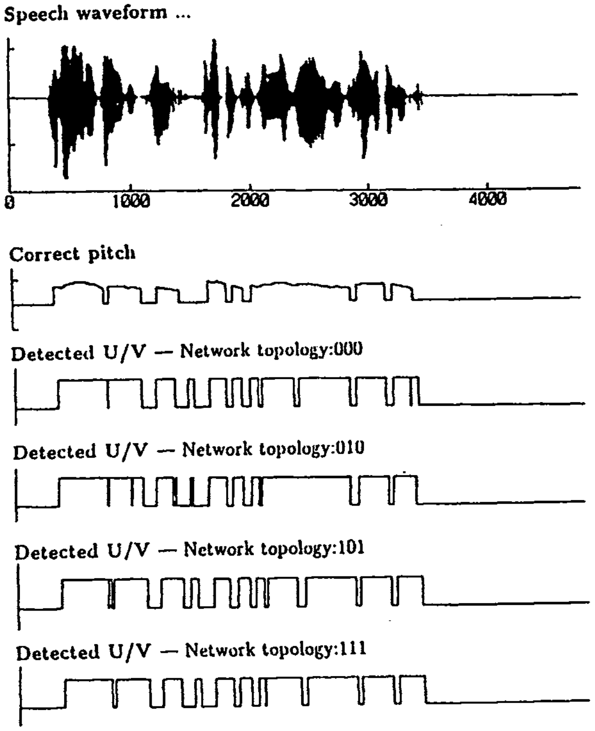


Fig. 3. An example of U/V detection for male speaker by using each NN trained with data 2 (t03-a01, non-training speaker).

It follows from the above, that the NNs with feedback coupling from the output layer to a hidden layer learn in particular the continuity of the time axis direction of time series patterns such as the U/V information, and it is effective in reducing the chattering of the U/V information in the boundaries where it is easily generated. However, the learning effect of time-continuity does not contribute directly to the improvement of the correct rate [13].

5.3. Pitch extraction and U/V detection results

Table 5 shows the pitch extraction and the U/V detection results using each NN trained with both data sets 1 and 2. We can see from this table that the net topology 111, i.e., CCNN-F, behaves the best for both pitch extraction and U/V detection. As for the former, there is an increase of about 3.4% in the correct rate and a decrease of about 0.3% in the percentage of GPE when compared with net topology 000; as for the latter, there is an increase of about 0.4% in the correct rate and a decrease of approximately 0.4% in the percentage of both VUE and UVE.

Now, the improvement of the correct rates in the pitch extraction NN of topology 101 and the U/V detection NN of topology 101 with respect to the net topology 000 are due to the larger number of the jointed training data. In the case of the CCNN-F, however, the highest improved accuracy can be obtained constantly, independent of the number of training data [13].

5.4. Comparison with other methods

Table 6 shows the U/V detection and the pitch extraction results using the NNs (net topology 000) trained with both data sets 1 and 2, compared with the results using the cepstral method and the LPC residual correlation methods [11]. For this case, we have not done post-processing such as the pitch smoothing technique, and we have compared directly the extracted values for each frame.

We see from this table that the correct rate and the error rates (UVE + VUE, GPE) are both better by the proposed NN

Table 3. Learning effects of time-continuity of U/V for each NN (non-training data). Average/variance: the U/V detection results with respect to the average/variance of the change from the preceding frames, divided by the correct values (increasing rate)

Net Topology	Training data 1NN				Training data 2NN			
	Well-known speakers		Unknown speakers		Well-known speakers		Unknown speakers	
	Average	Variance	Average	Variance	Average	Variance	Average	Variance
000	1.41	1.38	1.25	1.24	1.37	1.35	1.27	1.25
010	1.21	1.20	1.21	1.20	1.34	1.32	1.29	1.27
101	1.07	1.07	1.10	1.09	1.11	1.11	1.15	1.14
111	1.07	1.07	1.14	1.13	1.14	1.13	1.17	1.16

than by the other methods. By referring to the U/V detection results, when we consider an overall system that extracts the pitch only from well-detected voicing frames, the correct rate of the overall system is increased by more than about 4.4% and the error rate (UVE + VUE + GPE) is decreased by more than about 4.6% with respect to the other two methods, clearly showing the superiority of the proposed NN method. [The correct rate of the overall system is calculated here by taking as a correct frame one that has neither a UVE, a VUE, nor a pitch extraction error deviating from a range under ±5% correct pitch, and taking the percentage of these correct frames with respect to the total number of frames.] In the cepstral and the LPC residual correlation methods, the results are greatly influenced by the selection of the threshold value for the U/V detection; this value was determined here so as to make the correct rate of the overall system as large as possible.

Furthermore, comparing the CCNN-F method (net topology: 111) with the simple conventional NN method, there is a 2% increase to the 94.9% of correct rate in the overall system, and approximately 0.5% decrease to 3.8% of the error rate. Therefore we may say that the proposed CCNN-F model provides further improvement.

6. Conclusion

We carried out experimental comparisons for pitch extraction and U/V detection abilities based on different network structures, using BPFP banks and four types of multi-layered NNs. The following conclusions were obtained from the experimental results.

- The method using BPFP banks and conventional multi-layered NNs is a very effective pitch extraction method as compared to other typical techniques such as the cepstral method and the LPC residual correlation method.
- The CCNN-F structure with feedback couplings between the output and hidden layers, and cross couplings within the hidden layer is robust and produced the highest improved accuracy. Compared with the conventional multi-layered NN, it increases the correct rate by 2% and decreases the error (UVE + VUE + GPE) by approximately 0.5%; this is the synergistic effect of the feedback couplings and the cross-couplings.

Table 4. Comparison of correct rate of pitch extraction and U/V detection by using NN with feedback architecture (non-training data, a unit: %). The correct rate values of neural net topologies 101 and 111, relative values to these from neural net topologies 000 and 010 respectively

		Training data 1NN		Training data 2NN	
	Net topology	Well-known speakers	Unknown speakers	Well-known speakers	Unknown speakers
Pitch extraction	101	+0.52	+0.63	−0.21	−0.04
	111	−0.45	−0.55	+0.12	+0.98
U/V detection	101	+0.29	+0.69	+0.39	−0.26
	111	−0.53	+0.51	+0.57	+0.25

Table 5. Results of pitch extraction and U/V detection by using each NN with both data sets 1 and 2 (non-training, a unit: %)

Net topology	Pitch extraction		U/V detection		
	Correct rate	GPE	Correct rate	UVE	VUE
000	93.26	0.56	95.81	4.59	3.89
010	93.93	0.30	96.17	3.19	4.32
101	95.68	0.52	95.51	5.97	3.37
111	96.63	0.28	96.24	4.18	3.45

Table 6. Comparison with other methods of U/V detection and pitch extraction (non-training, a unit: %)

	U/V detection			Pitch extraction		Overall system		
Method	Correct rate	UVE	VUE	Correct rate	GPE	Correct rate	UVE + VUE + GPE	Remarks
cepstrum	86.12	7.69	18.58	78.15	1.07	84.26	14.48	Threshold = 2.0
LPC residual	93.63	11.26	2.66	88.36	4.48	88.52	8.92	Threshold = 0.15
NN	95.81	4.59	3.89	90.98	0.26	92.90	4.34	Net topology: 000
CCNN-F	96.24	4.18	3.45	94.19	0.08	94.90	3.81	Net topology: 111

- Training of the NN with cross-coupled hidden layers improves the accuracy of the discrimination function effectively in pitch extraction.
- The NN with feedback coupling from the output layer to a hidden layer learns in particular the time-continuity of the U/V information.

For future research, it is necessary to confirm the efficiency of the proposed CCNN-F model by carrying out listening experiments with the synthesized speech having as a voice source the extracted pitch and U/V information by the present method, and pitch extraction experiments in low-grade speech with noise. Also, it is necessary to pursue the relationship between the increment in the number of cross-coupling paths in a hidden layer and the accuracy improvement of the discrimination function. Furthermore, since it is not enough to learn the discrimination function, it is necessary to improve the approximate learning algorithm considering only the back-propagation of the previous time step of direct paths in the hidden layer.

REFERENCES

1. S. Furui. Acoustic and Speech Engineering. Kindai Kagaku-Sya Co. (1992). (in Japanese)
2. E. Barnard, R.A. Cole, M.P. Vea, and F..A. Allcva. Pitch detection with a neural-net classifier. IEEE Trans., Signal Processing, **39**, No. 2, pp. 298–307 (Feb. 1991).
3. H. Martinez-Alfaro and J.L. Contreras-Vidal. A robust real-time pitch detector based on neural networks. Proc. Int. Conf. ASSP, Toronto, Canada, pp. 521–523 (1991).
4. T. Ghiselli-Crippa and A. El-Jaroudi. A fast neural net training algorithm and its application to voiced-unvoiced-silence classification of speech. Proc. Int. Conf. ASSP, Toronto, Canada, pp. 441–444 (1991).
5. A. Ogihara and K. Fukunaga. A correcting method for pitch extraction using neural networks. I.E.I.C.E., Trans. Electron. **E77-A**, No. 6, pp. 1015–1022 (June 1994).
6. H. Miyabashi and T. Funada. Structure-dependency of performance for U/V detection NN. Tech. Rep. I.E.I.C.E., **SP94-107** (Jan. 1995).· (in Japanese)
7. T. Funada and T. Suzuki. A method for extracting pitch of speech signals by a bank of bandpass filter-pairs. Trans. I.E.I.C.E. (A), **J72**, No. 3, pp. 466–474 (March 1989). (in Japanese)
8. D.E. Rumelhart and J.L. McClelland (eds.). Parallel Distributed Processing. MIT Press (1988).
9. K. Doya and S. Yoshizawa. Adaptive neural oscillator using continuous-time back-propagation learning. Neural Networks, **2**, pp. 375–385 (1989).
10. K. Doya. Memorizing oscillatory patterns in neural networks. Computrol, **29**, pp. 52–62 (1990). (in Japanese)
11. J.D. Markel and A.H. Gray, Jr. Linear Prediction of Speech, Springer-Verlag, Berlin (1976).
12. T. Miura, ed. A new version: Auditory sensation and speech. I.E.I.C.E. (1980). (in Japanese)
13. H. Miyabayashi and T. Funada. Study on the effect of feedback loops and cross-coupled hidden layers in NN for pitch extraction. Tech. Rep. I.E.I.C.E., **SP95-41** (July 1995). (in Japanese)

AUTHORS (from left to right)

Hideo Miyabayashi received the B.E. and M.E. degrees in electrical engineering from Toyama Univ. in 1969 and 1971, respectively. He joined Japan Metals and Chemicals Co., in 1971, and since 1989 is an associate professor in Computer Engineering at Toyama National College of Maritime Technology. His research interests include neural networks, speech information processing, and parallel processing. He is a member of the Information Processing Society of Japan and the Society of Instrument and Control Engineers.

Tetsuo Funada graduated from Kanazawa University in 1966, and earned the M.E. and Dr.Eng. degrees in electronics engineering from Nagoya University in 1968 and 1971, respectively. In 1971, he joined the Faculty of Engineering at Kanazawa Univ. as an assistant professor, where he is currently a professor. His research interests include biological signal processing and speech information processing. He is a co-author of books such as *Foundations of Information Science* and *Foundations of Numerical Analysis*. He is a member of IEEE, the Acoustic Society of Japan, Japan Society of Medical Electronics and Biological Engineering and the Information Processing Society of Japan.