THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.28, No.2E 2009. 6. pp. 43~50

Prosodic Contour Generation for Korean Text-To-Speech System Using Artificial Neural Networks

Un-Cheon Lim*
*Dept. of Electronics Eng., Hoseo Univ.

(Received June 4 2009; accepted June 9 2009)

Abstract

To get more natural synthetic speech generated by a Korean TTS (Text-To-Speech) system, we have to know all the possible prosodic rules in Korean spoken language. We should find out these rules from linguistic, phonetic information or from real speech. In general, all of these rules should be integrated into a prosody-generation algorithm in a TTS system. But this algorithm cannot cover up all the possible prosodic rules in a language and it is not perfect, so the naturalness of synthesized speech cannot be as good as we expect. ANNs (Artificial Neural Networks) can be trained to learn the prosodic rules in Korean spoken language. To train and test ANNs, we need to prepare the prosodic patterns of all the phonemic segments in a prosodic corpus. A prosodic corpus will include meaningful sentences to represent all the possible prosodic rules. Sentences in the corpus were made by picking up a series of words from the list of PB (Phonetically Balanced) isolated words. These sentences in the corpus were read by speakers, recorded, and collected as a speech database. By analyzing recorded real speech, we can extract prosodic pattern about each phoneme, and assign them as target and test patterns for ANNs. ANNs can learn the prosody from natural speech and generate prosodic patterns of the central phonemic segment in phoneme strings as output response of ANNs when phoneme strings of a sentence are given to ANNs as input stimuli.

Keywords: *Korean TTS, Prosody, ANNs, Prosodic Corpus*

I. Introduction

Since mid 1970s, digital speech processing has been considered as one of the most convenient tools for man-machine interface [1]. In general, synthesis-by-rule and synthesis-by-concatenation methods have been used in TTS systems. A set of accurate prosodic rules has to be implemented in both methods to improve the naturalness of synthetic speech [2].

There are so many factors that affect the prosody in a sentence. They vary from emotional level to segmental level. At emotional level, there can be so many variable conditions, so we have to restrict to some specific conditions. At syntactic level, the style of a sentence, the number of syllables in a sentence or in a word and the accent in a word can affect the prosodic features of a phonemic segment in a sentence or in a prosodic phrase.

These prosodic rules can be extracted from syntactic theory or statistically from natural speech. But these rules are not perfect and inaccurate to exactly cover up all the possible prosodic rules of natural speech. So the algorithm for prosody generation in a TTS system has to be modified continually to adopt newly found prosodic rules or fix the wrongfully implemented prosodic rules to get more natural synthetic speech.

ANNs can learn the prosodic rules from real speech and generate them for a TTS system. They can

Corresponding author: Un-Cheon Lim (uclim@hoseo.edu)
Dept. of Electronics Eng., Hoseo Univ. 165, Sechul-Ri, Baebang-Myun, Asan-Si, Chungnam-Do, Korea. 336-795

learn the prosodic rules from natural speech if we could design the architecture of them properly and train them with sufficient prosodic information, and then there is no need to re-design an algorithm for new or fixed prosodic rules. We just need more prosodic data from natural speech to learn the prosodic rules in a prosodic phrase.

To train and test ANNs, a prosodic corpus which consist of a set of meaningful sentences and prosodic phrases should be made. Speech database spoken by speakers can be collected based on this corpus. Using analyzed parameters that have been estimated by short-time autocorrelation method, a sentence or a prosodic phrase can be divided into each phonemic segments. From these data, prosodic patterns for training and testing ANNs can be extracted. By applying curve fitting method to these prosodic contour of each phonemic segment, target and test patterns for ANNs can be estimated. ANNs can be trained with target patterns, and tested with test patterns.

II. Prosody in Korean Spoken Language

When we look at the parametric features of same kind of phonemes in a sentence, we can find out that their duration and energy and pitch variation - if they are voiced - are different from each other according to their position in a sentence or in a prosodic phrase. We call these variations of duration, intensity and pitch of phonemic segments in a sentence, in a phrase, or in a word according to various environmental conditions, the prosody. These parameters of each phonemic segment are supra-segmental features of time, not restricted to a single segment and they will vary at from acoustic level to emotional level. At emotional level, the prosodic parameters will be varied by speaker's affect, intent, personality and speaking rate. It's not easy to make prosodic rules that take account of all these variations, so we have to make some restrictions such as that a speaker should read the corpus several times at normal speaking rate and with flat reading style at emotional level and real speech spoken by the speaker can be recorded and collected as speech database for prosody [3-7].

2.1. Duration of a phonemic segment

The factors that affect the length (duration) of a phonemic segment in a prosodic phrase are preceding and following phonemes, if it is stressed or not, if there is a stop or a pause, the number of phonemes in a prosodic phrase, a relative frequency of the word that includes this phoneme, etc.

There were some studies that said the length of a vowel with following unvoiced consonant is shorter than that with voiced one and the duration of a vowel preceding a fricative consonant is longer than that stop consonant etc. As the number of syllables in a word increases, the duration of each segment should be decreased with different rate within a certain limit.

We can see the length and energy contour of each voiced segment will vary from a speech sample of a Korean sentence shown in Fig. 1.

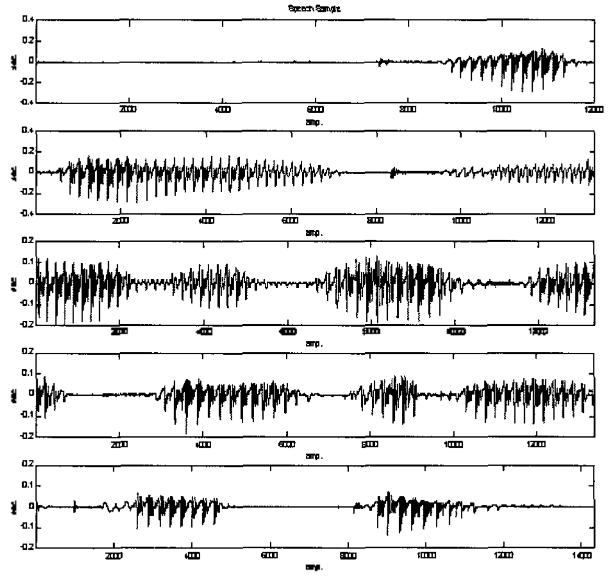


Fig. 1. Speech Sample of a Korean sentence.

2.2. Pitch contour of a phonemic segment

In pitch controlled languages, the pitch has been considered as an indicator of the syntactic structure of sentences read, so most of TTS systems use the concept of a baseline in fundamental frequency (pitch) and implement it in a sentence or in a phrase [8] [9].

In Korean spoken language, syntactic boundaries such as phrase boundaries, accents in word, and segmental phonemes are said to affect the pitch period variations of each phoneme in a sentence.

It's pointed out that a baseline of pitch can be detected on a prosodic phrase rather than on a sentence and the relative position of a phoneme in a prosodic phrase is more important than in a sentence.

There are papers said that prosodic cues play an active role not only at syntactic level, but also at lexical and syllabic level as well. The lexical stress pattern in a word play an important role, and the sentential level stress is more important. There are also many studies about the pitch, duration and amplitude variation in each segment as a result of the interaction between successive segments [10] [11].

The number of phonemes in a sentence can be so large, a speaker can not spell that sentence without any pause. So it's more reasonable to think about that prosody in a prosodic phrase rather than in a sentence should be learned.

There are some papers said that the prosodic boundaries in Korean spoken language can be detected, so a sentence can be divided to a number of prosodic phrases. The strengths of prosodic boundaries also can be detected by acoustical and perceptual features. We should know that prosodic features of a phoneme in a sentence can be varied according to the environmental conditions around that phoneme.

If we consider the starting phoneme of a prosodic phrase acts as if the starting phoneme of a sentence, it will be more efficient to learn the prosodic features of prosodic phrase rather than those of a sentence.

In Fig. 2, the pitch contour of a phoneme /yae/ and it's regression line estimated by 2nd order polynomial equation are shown. The duration and the pitch contour of same phoneme from different conditions will not be same as in Fig. 2.

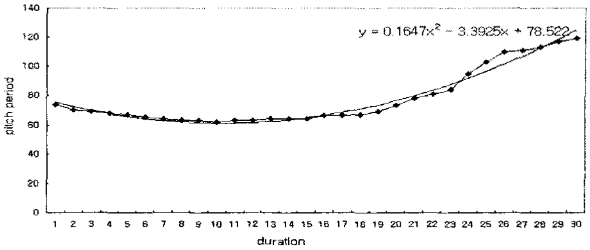


Fig. 2. The pitch contour of a phoneme /yae/ and it' s regressive line.

2.3. Energy contour of a phonemic segment

There are few papers about the intensity variation of phonemes in a sentence. But in Korean the intensity of one phoneme is varying according to it's own sonority, the stress pattern in a word, and the relative position in a sentence. There are some papers said that prosodic cues play an active role not only at syntactic level, but also at lexical and syllabic level as well. The lexical stress pattern in a word plays an important role, and the sentential level stress is more important.

The energy contour of a phoneme in Fig. 2 and it's regression line are shown in Fig. 3. As in Fig. 2, the length and the energy contour of same phoneme from different conditions will vary according to the prosodic rules.

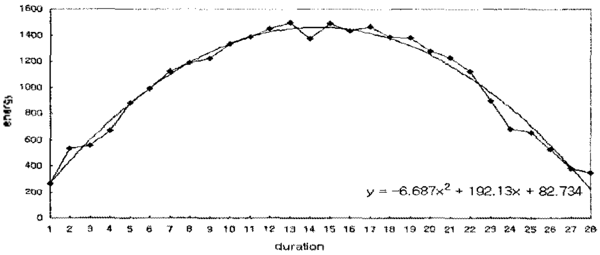


Fig. 3. The energy contour of a phoneme /yae/ and it's regressive line.

2.4. Prosodic contour of a phonemic segment

We have to integrate all these prosodic rules into a prosodic algorithm when a TTS system adopt an algorithmic approach for prosody generation. But as you can see, there are so many variations which we have to take account of and all of the prosodic rules cannot be integrated into one prosodic algorithm. Futhermore if there is a need to add newly found prosodic rules and fix wrongfully implemented prosodic rules, we have to re-design it and program it again.

If there is a way to learn all these rules about prosody from real speech, and there is no need to modify them again and again, it will be a good approach. ANNs can be a good candidate for prosody generation without any re-design of the prosodic algorithm [12-16].

To make ANNs learn the prosodic rules residing in natural speech, we have to design the architecture of ANNs in advance. We have to take account of the environmental conditions of a phonemic segment to learn and generate it's prosodic information accurately [17-19].

We can find out the pitch and energy contours of each voiced phoneme and the energy contour of each unvoiced phoneme in a prosodic phrase.

The spectral and prosodic variations of phonemes can be extracted through short-time analysis on real speech. If the sampling frequency is 10 KHz, the size of short-time window is 256 samples, and the overlapping interval is 128 samples, the time interval per one window frame will be 12.8 msec. The number of phonemes varies from 2 to 11 in almost all of the prosodic phrases. There is a case that the number of phonemes is 24, but it's very rare.

We can approximate the pitch and energy contour for each phonemic segment using nonlinear curve fitting method. To do that, we have to accurately segment each phoneme within analyzed data and determine it's duration and the contours of pitch and energy.

The pitch and energy contour of one phonemic segment can be approximated with polynomial equations. The order of polynomial equation can be set to 2 or 3.

We can approximate the pitch contour of a phonemic segment in a prosodic phrase with 2nd order polynomial equation (1), and the energy contour, with equation (2)

$$p(n) = p_2 \times n^2 + p_{1\times}n + p_0, 0 \le n \le d-1 \qquad (1)$$

$$e(n) = e_2 \times n^2 + e_{1\times}n + e_0, 0 \le n \le d-1 \qquad (2)$$

where d is the number of frames of a phonemic segment, p_2, p_1 are polynomial parameters of it's pitch contour, and p_0 is an initial pitch period, e_2, e_1 are polynomial coefficients of it's energy contour, and e_0 is it's initial energy.

We can make ANNs learn these contours of each phonemic segment in a prosodic phrase, as we apply those prosodic patterns to the input layer of ANNs.

III. ANNs for prosody generation

ANNs are simplified models of the central nervous system that have the ability to respond the input stimuli and to learn to adapt to the environment. These ANNs have the special features such as robust performance when dealing with noisy or incomplete input patterns, a high degree of fault tolerance, high parallel computation rates, the ability to generalize and adaptive learning, so they can be used as effective computational processors for various tasks including pattern recognition, classification, data compression and modeling and forecasting, etc.

There are many kinds of ANNs according to the learning paradigm, the network architecture and general area of application. To learn or predict the prosodic contour of a phonemic segment, ANNs such as Multi Layer Feed Forward (MLFF) networks with Back Propagation (BP) can be used.

If BP ANNs can learn prosodic rules from real speech, there will be no need to re-design a prosody algorithm to include new prosodic rules and to fix wrong rules which will degrade synthetic speech quality, and they can also learn prosodic rules in Korean sentences that can not be easily defined in an algorithmic approach.

We can make just one ANN to generate all the prosodic information about a phoneme in a sentence. To do this, the structure of the ANN will become too complex and the computation time will be too huge.

So we can use 2 BP ANNs which will learn the pitch or energy contour of phonemic segment in Korean spoken language.

We can think of a simple ANN that will get a phoneme as an input and generate it's prosodic information as an output. But it is too simple to learn supra-segmental effects of it's surrounding phonemes, effects of it's relative position in a prosodic phrase and of syntactic symbols near it. In order to take account of the supra-segmental effects of the preceding and following phonemes and the lengthening of the final syllable in a prosodic phrase and the effects according to it's relative position in a prosodic phrase, we can assign 11 phoneme strings as an input stimuli to each ANN. From these phoneme strings, the 6th phoneme will be the central phoneme, and prosodic features of this central phoneme will be output parameters of each ANN.

The number of phonemes in a prosodic phrase needs not to be restricted to a small number, but we can train ANNs with a small number of input nodes more efficiently.

The input layer of each ANN consists of 11 nodes that represent a set of 11 phoneme strings that includes symbol such as a dot, a comma, or a blank.

For nonlinear mapping between the input and output pattern, we have to adopt hidden layers. We can select one hidden layer architecture for the learning of prosodic information, and let the number of nodes of hidden layer be the same as the number of input nodes.

The output layer of pitch ANN will be consisted of 4 modules which can generate a duration and 3 polynomial parameters for the pitch contour of central phoneme from a set of 11 phoneme strings.

A typical block diagram of BP ANN with one hidden layer for pitch contour generation is shown in Fig. 4. This BP ANN can be trained to learn pitch variation of central phoneme of input phoneme strings.

We can design energy BP ANN for energy contour generation that has the same architecture as the pitch BP ANN.

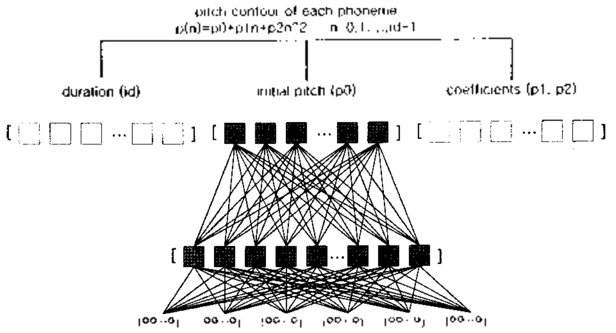

pitch contour of each phoneme
p(n)=p0+p1n+p2n^2 n 0,1, ..,id-1
duration (id)
initial pitch (p0)
coefficients (p1, p2)

Fig. 4. Block diagram of BP ANN for pitch contour generation.

The input layer of BP ANN will take a string of 11 phonemes as input stimuli, and the output layer of it will learn and generate the prosodic parameters of the central phoneme of the input phoneme strings.

We can find out 18 initial consonants, 21 medial vowels and 7 final consonants available in a syllable after going through a proper phonological processing in a Korean sentence. In addition to these phonemes, we have to include some syntactic symbols like period, comma, question mark and blank that will be source of supra-segmental effects in prosodic variations.

The output layer of pitch BP ANN will output the duration of central phoneme and each polynomial parameters of the pitch contour of the central phoneme.

The duration of central phoneme will be the number of frames of that phoneme, which we can be determined from the analyzed speech data by segmentation. It will not exceed 24 frames.

To prepare target and test patterns for each phoneme in a prosodic phrase, the speech database spoken by a speaker should be recorded and collected and analyzed with short-time analysis algorithm, and then the prosodic pattern of pitch and energy contour about each phoneme and the duration of that can be made for the training and testing BP ANNs.

The pitch and energy contour of a phonemic segment can be approximated with 2nd order polynomial equation. So the target and test patterns for BP ANNs will be consisted of a duration (d), 6 polynomial parameters (p_0, p_1, p_2, e_0, e_1, e_2) of it.

In learning phase, the BP ANNs will be trained to learn the prosodic rules when a set of phoneme strings is assigned to the input layer of BP ANNs and the prosodic patterns of it's central phoneme are given to the output layer of BP ANNs.

In BP ANNs, a sigmoid function can be used as a transfer function of an artificial neuron unit. To minimize the total error between the output pattern and target pattern, BP ANNs adopt gradient descent learning method to adjust weights in networks. When the phoneme strings are given to ANNs, they generate an output pattern of the central phoneme, we compare it with a target pattern by computing Root Mean Square (RMS) error between output and target pattern.

Let the input pattern of the network, **x**, the weighting matrix between input layer and hidden layer, V, the input to the hidden unit **H**, the output of hidden unit, **y**, the weighting matrix between hidden layer and output layer, W, the input to the output node, **I**, the actual output of output node, **z**, and the target pattern, **t**. Then the total error E_{total} can be defined in equation (3).

$$E_{total} = E_{total} + 1/2 \,||t - z||^2 \qquad (3)$$

where $z = F(I)$, $I = Wy$, $y = F(H)$, $H = Vx$.

To minimize this RMS error E_{total}, the networks have to adjust output layer weights W and hidden layer weights V according to the BP learning method.

To adjust the weights, error signal vectors δ_{zk}, δ_{yj} should be computed as in equation (4) and (5).

$$\delta_{zk} = 1/2\ (t_k - z_k)\ (1 - z_k^2),\ \text{for } k = 1,..,K \qquad (4)$$

$$\delta_{yj} = 1/2\ (1 - y_j^2)\Sigma\delta_{zk}w_{kj},\ \text{for } j = 1,..,M \qquad (5)$$

Then the output layer weights are adjusted according to the equation (6), and the hidden layer weights are adjusted according to the equation (7).

$$w_{kj}^{new} = w_{kj}^{old} + \eta\delta zkyj,\ \text{for } k,\ j = 1,..,K \qquad (6)$$

$$v_{ji}^{new} = v_{ji}^{old} + \eta\delta_{yj}x_i,\ \text{for } j = 1,..,M \qquad (7)$$

$$I = 1,..,M$$

where K is the number of output units, the M is the number of hidden and input units and η is predetermined learning rate.

We call this training cycle one epoch. We set the maximum number of epochs to 200, and if the RMS error goes below the error threshold, then let the networks stop the weight adjustment, start new adjustment for the patterns of new phoneme.

When the training for a phoneme was done within a pre-determined tolerance, we have to shift the phoneme strings to the left, so the next phoneme will become the central phoneme of shifted phoneme strings.

The performance of these BP ANNs in learning phase can be estimated by counting the number of phonemes that can't be learned with an RMS error below the threshold within maximum epochs.

After all the target patterns have been trained, the performance of each BP ANNs could be tested with test patterns.

IV. Training and testing ANNs

To train and test BP ANNs for prosody generation of phonemic segment in a Korean prosodic phrase, a corpus for prosody should be made according to certain guidelines. Natural speech repeatedly spoken by special speakers based on the prosodic corpus should be recorded and collected as speech database. By analyzing these speech data, the parametric representation will be collected. By segmenting each each phonemic segment based on the parametric variations, the prosodic features of each phonemic segment can be acquired. The prosodic patterns for each phonemic segment can be estimated by 2nd order polynomial equations. The duration and the polynomial coefficients of each phonemic segment will be used as the target patterns in the output layer of BP ANNs.

4.1. Prosodic corpus

A prosodic corpus consists of some meaningful sentences or phrases, which are made by the words selected from the list of 412 PB words.

4.2. Speech database

Each sentence or phrase in a prosodic corpus is read by special speaker multiple times in a row and will be recorded and be collected as speech database.

4.3. Analysis and Segmentation

The recorded speech data were analyzed with short-time autocorrelation algorithm. The segmentation of each phonemic segment will be made according to the parametric variations. Multiple sets of prosodic patterns should be prepared for the training and testing ANNs.

4.4. Prosodic patterns for BP ANNs

The pitch and energy contour of each phonemic segment in a prosodic phrase will be approximated to the polynomial coefficients by curve fitting method, and the polynomial coefficients of each phonemic segment and it's duration will be made into a pitch target pattern for pitch BP ANN and a energy target pattern for energy BP ANN.

4.5. Training ANNs

In training phase, some sets of prosodic patterns should be used as target patterns in the output layer of ANNs.

The 11 phoneme strings in a sentence and the target pattern of the central phoneme (6th phoneme in the strings) will be given to the BP ANNs, and the total RMS error of each network should be computed. To minimize this total system error, the adjustment of the weights of output and hidden layer will be computed.

The BP ANNs will be trained within pre-determined maximum epochs, and when the total system error goes down the threshold values for the minimum error, new training cycle for a prosodic pattern of another phonemic segment will be started without any further training for the previous segment to the maximum epochs.

The performance of each BP ANN can be computed in training phase.

4.6. Testing ANNs

In test phase, the output patterns of each BP ANN will be compared to the test prosodic patterns reserved for the test, and the similarity between these two patterns will be computed without any training.

V. Conclusion and discussion

In the experiment of 3 sets of prosodic patterns, the performances of each BP ANN were near about 91% in training phase, and 89% in test phase [19]. It shows that ANNs can learn the prosodic features from real speech.

Because the number of phoneme strings was set to 11, the supra-segmental effect and the effect of accent in a word can be covered, but if the number of phoneme strings in a prosodic phrase will exceed 11, then the ANNs can not fully learn the effect of phoneme's relative position in a prosodic phrase. To learn these supra-segmental effects, we have to increase the number of units in the input layer of ANNs in spite of huge computational load.

If we can increase the number of sentences in a prosodic corpus, increase the number of repeats for the same sentences, train ANNs with multiple sets of target patterns, we don't have to be trapped into over-training and can improve the system performance and get more natural synthetic speech.

If we can integrate this prosody generation system into a TTS system and get subjective score like Mean Opinion Score about the naturalness of the synthetic speech through listening test, then we can see the adaptability of ANNs.

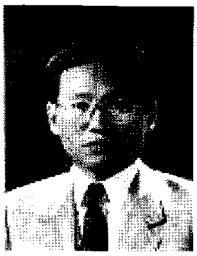
References

1. J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*, Springer-Verlag, 1976.

2. J. Allen, M. S. Hunnicutt and D. H. Klatt et al, *From Text To Speech*, Cambridge University Press, 1987.
3. A. Waibel, *Prosody and Speech Recognition*, Morgan Kaufmann Publishers, 1988.
4. A. M. Liberman et al, "Minimal rules for synthesizing speech," *J. Acoust. Soc. Am.*, vol. 31, no. 11, pp. 1490–1499, Nov. 1959.
5. J. Allen, "Synthesis of speech from unrestricted text," *Proc. IEEE*, vol. 64, no .4, pp. 433–442, Apr. 1976.
6. N. Umeda, "Vowel duration in American English," *J. Acoust. Soc. Am.*, vol. 56, pp. 434–445, 1975.
7. J. Pierrehumbert, "Synthesizing intonation," *J. Acoust. Soc. Am.*, vol. 70, no. 4, pp. 985–995, Oct. 1981.
8. R. M. Meli and F. Fallside, "The modeling of F0 contours," in *IEEE Proc. ICASSP* '82, pp. 947–949, 1982.
9. M. Ljungqvist and H. Fujisaki, "Generating Intonation for Swedish Text-to-Speech Conversion Using a Quantitative Model for the F0 Contour," in *Proc. Eurospeech '93*, pp. 873–876, 1993.
10. Hyun Bok Lee, "Korean prosody : Speech rhythm and intonation," *Korea Journal*, pp. 42–69, Feb. 1987.
11. J. C. Lee, S. H. Kim and M. Hahn, "Intonation Processing for Korean TTS Conversion Using Stylization Method," in *Proc. ICSPAT '95*, vol. II, pp. 1943–1946, 1995.
12. C. Tuerk and T. Robinson, "Speech Synthesis Using Artificial Neural Networks Trained on Cepstral Coefficients," in *Proc. Eurospeech '93*, pp. 1713–1716, 1993.
13. M. Riedi, "A Neural-Network-Based Model of Segmental Duration for Speech Synthesis," in *Proc. Eurospeech '95*, vol. I, pp. 599–602, 1995.
14. D. P. Morgan and C. L. Scofield, *Neural Networks and Speech Processing*, Kluwer Academic Pub., 1991.
15. Mazin G. Rahim, *Artificial Neural Networks for Speech Analysis/Synthesis*, Chapman & Hall, 1994.
16. Adam Blum, *Neural Networks in C++*, John Wiley & Sons Inc., 1992.
17. Sok Wang Chang, Hyun Joon Kim, Chang Su Ryoo, Un-Cheon Lim, "A Study on the Prosody Generation in Isolated Words with an Artificial Neural Network," in *Proc. ICSP'97*, vol. 1 of 2, pp. 207–210, 1997.
18. Kyung-Joong Min, Un-Cheon Lim, "Architecture of Artificial Neural Networks for Prosody Generation in Korean Sentences", *Proc. ICSP'2001*, vol. 2 of 2, pp. 771–776, 2001.
19. Kyoung-Joong Min, Un-Cheon Lim, "Korean Prosody Generation and Artificial Neural Networks", *INTERSPEECH 2004 ICSLP*, vol. 3 of 8, pp. 1869–1872, 2004.

[Profile]

- Un-Cheon Lim



1981.2: Seoul National Univ. Dept. of Electronic Eng. (B.E.)
1983.2: Seoul National Univ. Dept. of Electronic Eng. (M.E.)
1991.8: Seoul National Univ. Dept. of Electronic Eng. (Ph.D.)
1984.3~present: Hoseo Univ. Professor
※Main Research: Speech Analysis and Synthesis, Artificial Neural Networks