Energy Contour Generation for a Sentence Using a Neural Network Learning Method

Jungchul Lee, Donggyu Kang*, Sanghoon Kim*, Koengmo Sung***

* ETRI, 161 Kajong-Dong, Yusong-Gu, Taejon, 305-350, Korea

** Department of Electronic Engineering, Seoul National University

ABSTRACT

Energy contour in a sentence is one of major factors that affect the naturalness of synthetic speech. In this paper, we propose a method to control the energy contour for the enhancement in the naturalness of Korean synthetic speech. Our algorithm adopts syllable as a basic unit and predicts the peak amplitude for each syllable in a word using a neural network (NN). We utilize indirect linguistic features as well as acoustic features of phonemes as input data to the NN to accommodate the grammatical effects of words in a sentence. The simulation results show that prediction error is less than 10% and our algorithm is very effective for analysis/synthesis of the energy contour of a sentence, and generates a fairly good declarative contour for TTS.

1. INTRODUCTION

The functionality of TTS is to provide various information to the user through synthetic speech. A user can utilize TTS for reading unrestricted texts or other character data. For this purpose, TTS has to be able to produce high quality synthetic speech in the sense of intelligibility and naturalness. A speaker uses the phoneme duration, the pitch, and the energy and pause to represent his intention more clearly and naturally to the listener. So the prosodic phenomena of an utterance is the result of various factors such as semantics, syntactic structure , words, articulation, intention, and the speaking rate.

Many researchers have devoted a lot of efforts to improve synthetic speech naturalness. Due to the progresses in language processing technique, it becomes possible to predict the position and the duration of pauses in a sentence. Many research results successfully predict segmental duration and generate a proper pitch contour for the corresponding sentence. Concerning about the energy feature, there were some studies on the relation between the focus and intensity of speech [1]. Turk reported that the variation of segmental duration is more dominant factor to the prominence than the intensity through the perception test [2]. But this result is restricted to the prominence and dose not mean that energy contour is insignificant factor for naturalness. Kuniszyk-Jozkowiak suggested that energy pattern could be an effective factor to discriminate the fluency of utterance [3]. Recently, Hanson tried to control the amplitude of vowels in a sentence using some rules to improve the naturalness of synthetic speech [4]. But still there are little research results related to the energy contour control for a sentence.

Many TTS synthesizers adopt a segment concatenation technique, and the segment unit is usually diphone, demisyllable, or triphone. If these speech segments are selected from monosyllable or di-syllable speech database, the energy contour for each segment shows a simple pattern; rising at beginning and falling at the end. So simple segment concatenation makes an uniform syllabic energy pattern and results in an unnatural synthetic speech even though the segmental duration and the intonation are properly controlled. Even if the speech segments are collected from isolated words or sentence, the naturalness at a sentence level is still an open problem.

In this paper, we focused our study for the energy contour modeling to improve the naturalness of synthetic speech in the sense of syntactic structure, grammatical categories, and phonetic environment rather than semantics, intention, or speech rate. We observed that there are many regularities of energy contour patterns for sentences and a neural network can predict speech envelope for a sentence represented by syllabic peak values in a sentence [5-6].

Our algorithm consists of three steps. Firstly we predict the energy contour for each word in a sentence using a neural network. The inputs to the NN are acoustic features of phonemic environment, word position in a sentence, preceding pause duration, following pause duration, average pitch values of preceding, target, and following words. The outputs of NN are peak values of corresponding syllables in a word. In the next stage, the local gain is multiplied to the synthetic unit segment to adjust the peak value to the predicted value. Finally, the energy contour is smoothed at the voiced segment boundaries.

2. SPEECH MATERIALS AND ENVELOPE PATTERN ANALYSIS

The speech material for this study consists of read sentences. These sentences are chosen randomly from a large text corpus, and read by a professional native Korean male announcer, yielding 156 sentences containing 2,186 words, 6,632 syllables. Speech data are recorded using a DAT recorder and digitized at 16kHz with 16bit resolution. These speech data are hand-labeled with phoneme symbols, syllable, and word boundary markers.

Figure 1 depicts energy contours for 156 sentences where syllable is a basic unit to represent energy contour of a sentence. Horizontal axis denotes the position of a syllable in a sentence normalized with the total number of syllables in the sentence and vertical axis means the syllable energy that is defined as a peak amplitude in the syllable. The advantage of using syllabic peak value is to minimize local variation and effectively shows

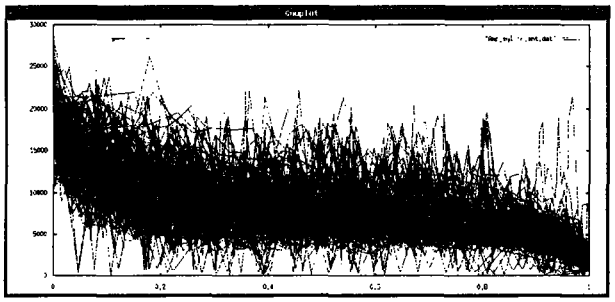

Figure 1: Energy contours for 156 sentences represented by syllabic energy values

global speech envelope. The overall energy contours for sentences show gradually decreasing pattern and some reset points in a sentence that can be deduced from energy values of syllables. The dynamic range of amplitude variation is also gradually decreased according to the position in a sentence. The energy contour for a sentence represented by peak amplitude of each word also shows a similar pattern.

We investigated the word energy contours according to the number of syllables in a word. Figure 2 illustrates word energy patterns that are represented by syllabic energy. To remove the positional effects, such as the global declination due to the position in a sentence, DC offset values are subtracted from each word. These figures show little correlation between the number of syllables and the word energy pattern or between the syllabic peak and the position of syllable in a word. And it can be also observed from this figure that the energy contour can not be determined by simple factors such as position of word or syllable. From these results, we can infer that we need more information for reliable energy prediction in addition to the word position in a sentence and number of syllables in a word.

From the analysis results of word energy contours, we found that phonetic environment for words and inherent energy

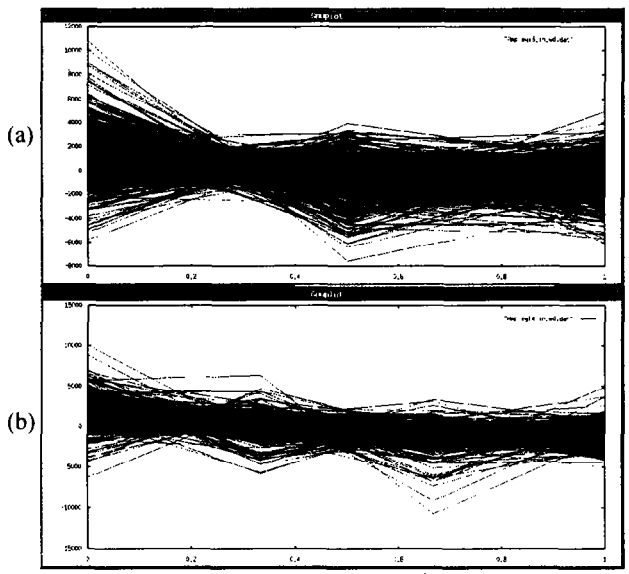

Figure 2: Energy contours of (a) 3 syllabic words, (b) 4 syllabic words.

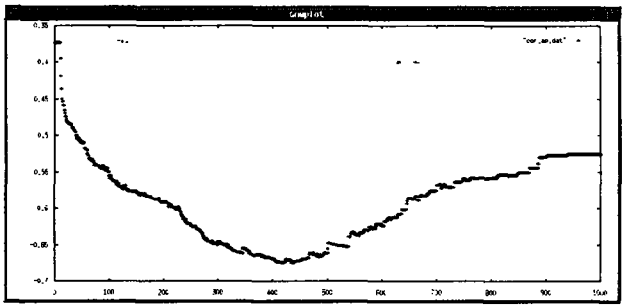

Figure 3: Correlation between the word energy and the position of word in a virtual breath group triggered by pause duration threshold [msec].

patterns are not sufficient to determine the energy contour for a specific word in a sentence. Figure 3 depicts the correlation between the word energy and the position of the word in a breath group. Horizontal axis in Figure 3 denotes the threshold value, P_{th}, of pause duration in msec to identify the breath group boundary. The correlation coefficient, r_{ap}, of word energy and the word position in a breath group can be calculated from

$$r_{ap} = \frac{E\left\{\left(A-\eta_a\right)\left(P-\eta_p\right)\right\}}{\sqrt{E\left\{\left(A-\eta_a\right)^2\right\}E\left\{\left(P-\eta_P\right)^2\right\}}}, \qquad (1)$$

where A is the word energy, η_a is the mean value of all word energy, P is the position in a breath group, η_p is the mean of the number of words in a breath group. When P_{th} is set to around 400msec, correlation value is negatively maximized to –0.68, and this means that it is reasonable to assume that the word energy is high for an initial word and is scaled downward within the breath group defined by the pause duration. As P_{th} becomes smaller or greater than the optimal value, the word energy value is less correlated with the position of word in a virtual breath group because there is little consistency in assigning a breath group boundary.

Prosodic features such as intonation, pause, segmental duration, and energy are highly dependent upon syntactic structure, part of speech, and grammatical function information. This means those features in speech can be assumed to be highly correlated to each other. Hence we can use other prosodic features as indirect factors instead of directly utilizing linguistic information in analyzing or generating the energy contour. Especially for TTS, it is more efficient to utilize the pause information and pitch contour in calculating energy contour because TTS generally uses linguistic information to predict pause and pitch values. To verify the correlation between the energy and other prosodic features, we investigated our speech data.

From the word energy distribution due to the pause duration preceding a target word, we can find a general tendency of the proportional relation between the word energy and the preceding pause duration. Correlation coefficient value is 0.7 for preceding pause and a target word energy. But the range of energy variation due to the pause duration is large because the

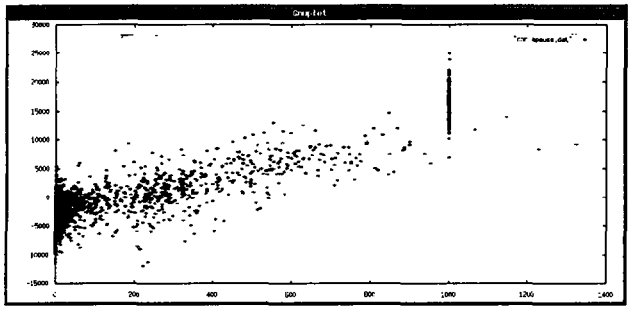

Figure 4: Distribution of energy difference between the two adjacent words due to the intervening pause.

effect of word position is not considered.

Figure 4 shows the distribution of energy difference between the two adjacent words according to the duration of intervening pause where horizontal axis and vertical axis denote the pause duration and the energy gap, respectively. In this figure, the effect of word position in sentence is minimized and we can find more clearly the proportional relation between the energy difference and the pause duration in the range of 200 msec and 600 msec. If the pause is bellow 200 msec, the energy falling occurs more frequently and this supports the assumption that if a short pause is inserted, the energy contour continues the declination pattern rather than reset. For the pauses above 600 msec duration, the pauses make almost the same effects on the energy difference.

Figure 5 shows the word energy distribution according to the word mean pitch value. Words having mean pitch values smaller than 80 Hz are mostly final words in intonational phrases, clauses or sentences and these words usually have low energy values. This figure shows the tendency that the word energy is in proportion to its word mean pitch value. The degree of correlation between the word energy and its mean pitch value is 0.7. From Figure5 showing that standard deviation is also proportional to the pitch value, we can infer that the word mean pitch value only is insufficient to predict the word energy and we need more information for reliable energy prediction.

We examined the contribution of the pitch difference between two adjacent words to the energy difference and find that if the pitch value of following word is greater than the preceding one, the energy of following word is increased proportionally. But

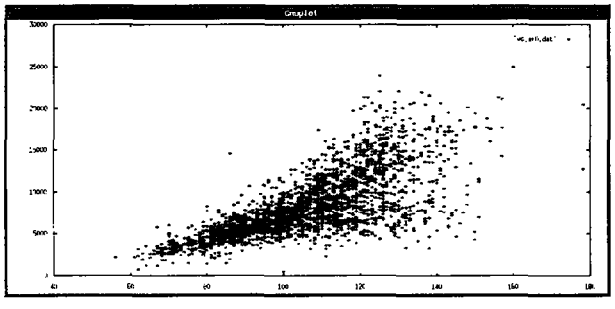

Figure 5: Word energy distribution according to the word mean pitch value

when the mean pitch value of a word is less than or almost same as that of the preceding word, the pitch difference is not an adequate factor for predicting the word energy.

From these analysis results, we can infer that the energy envelope for a word is a result of many factors, such as grammatical function of a word in syntactic structure, position in a sentence, common usage, and phonetic structure of a word. And we found that some of these factors can be replaced with prosodic information such as the pause and the pitch value to predict the energy pattern of a word.

3. MODEL FOR WORD ENERGY CONTOUR GENERATION

Our algorithm adopts syllable as a basic unit to represent the speech envelope, and the syllabic energy is defined as the peak amplitude in a syllable. The sentence energy contour generated through the three-step process. The first step predicts syllabic energy values in a word using the phonetic structure of a word, prosodic information, and the position of the word in a sentence. For the phonetic structure representation of a word, we use position of articulation, manner of articulation, intensity, and voiced/unvoiced information for each phoneme in a word. Phonetic structure information also includes the typical syllabic energy values to accommodate the effect of syllable structure type such as V, CV, CVC. Prosodic information means the duration of pauses before and after a target word, the mean pitch values of target word and adjacent two words. The second step multiplies the peak amplitude of each speech segment by a gain constant to fit to the predicted amplitude. This process does not change the envelope shape of each segment but multiplies a gain term only to modify the maximum peak value of a segment. In the last step, the energy contour is smoothed at the concatenation boundary of two adjacent voiced segments.

It is difficult to make energy control rules because of the complicate interaction between the factors affecting the energy contour of a word. So we adopt a three-layer neural network as shown in Figure 6 to predict the syllabic energy in a word. An NN has an advantage for modeling of non-linear relation even though it has disadvantages such as no accountability and difficulty in modification without retraining.

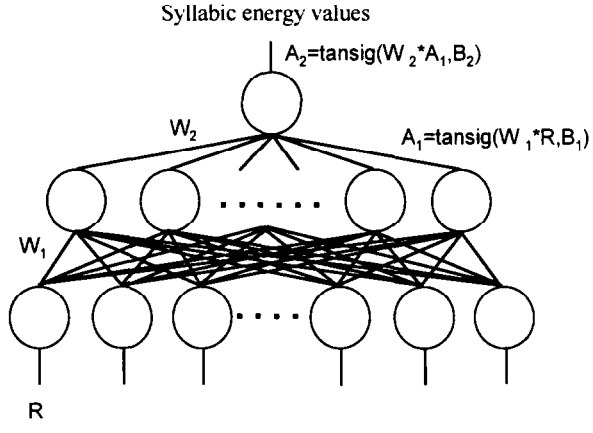

Figure 6: The structure of a word energy contour predictor.

The network is trained with words from the speech material and the inputs to the NN are phonetic, prosodic features and word position in a sentence, while the outputs are syllabic energy values for a target word. In the training process, the prosodic informations are extracted from real speech. In the prediction process of TTS, they are calculated by some rules and melodic tables.

4. EXPERIMENTAL RESULTS

We have performed two experiments. The speech material for our present study consists of recorded 156 sentences. Firstly, we trained the syllabic energy contour in a word with a neural network and Figure 7 shows our prediction result. These results came from the independent neural networks with respect to the number of syllables in a word. Hence our results can be a strong proof that the neural network can be very effective to predict the word energy contour, and to improve the naturalness of the synthesized speech. Table 1 shows the prediction error. These results came from the independent neural networks with respect to the number of syllables in a word. These results can be a strong proof that the neural network can be very effective to predict the word energy contour, and to improve the naturalness of the synthesized speech. Secondary, we implemented this algorithm in TTS and generated synthetic speech. Figure 8 shows both the original and the rule-generated energy contours for a test sentence. The mean absolute deviation between the original contour and the rule-generated contour was less than 10%. In other words, the rule-generated energy contour traces the original contour fairly well. In perception test, the synthesized speech with the synthesized energy contour is much similar to that with the original energy contour.

No. of syllables	$E\{\lvert E_{target}-E_{predict}\rvert / E_{target}\}$
1	4.6%
2	7.2%
3	9.6%
4	8.2%
5	3.9%
6	3.6%
7	0.2%
8	0.2%
9	2.5%

Table 1: Standard deviation of the prediction error according to the number of syllables in a word.

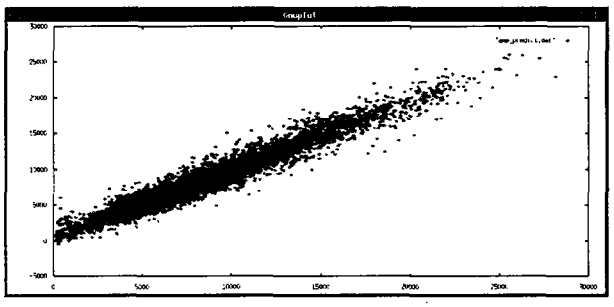

Figure 7: Target energy value and prediction values using a NN for the syllabic energy

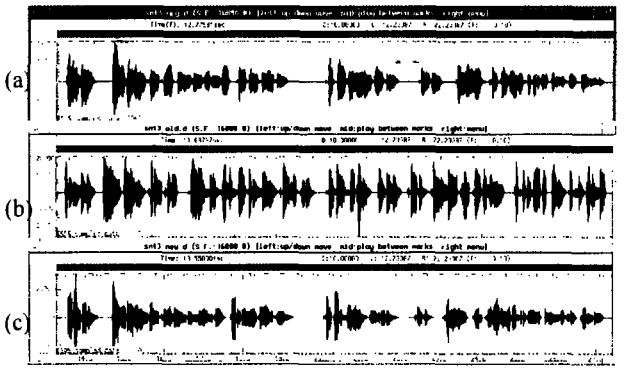

Figure 8: (a) Original, (b) synthetic speech without energy contour control, (c) synthetic speech with energy contour control

5. CONCLUSION

In this paper, we proposed a new model for synthesizing energy contours for sentences by utilizing a neural network learning method. The simulation results show that our algorithm is very effective for analysis/synthesis of the sentence energy contour while generating a fairly good declarative contour. Our future works include the improvement of the prediction accuracy with a large speech database. Also, we will try to expand this model to accommodate dialogue-type speech.

6. ACKNOWLEDGEMENT

This work is supported by the Korea Ministry of Information Communication under the research project of "Speech Input/ Output Processing Technology for Human-Computer Interface".

7. REFERENCES

1. D.Hirst, "Prediction of prosody: An overview," in Talking Machines:Theories, Models, and Designs, North-Holland, pp.199-204, 1992.
2. A.E. Turk, J.R. Sawusch, "The processing of duration and intensity cues to prominence," J. Acoust. Soc. Am. 99, pp.3782-3790, 1996.
3. W. Kuniszyk-Jozkowiak, "A comparison of speech envelopes of stutterers and nonstutterers," J. Acoust. Soc. Am. 100, pp.1105-1110, 1996.
4. H.M. Hanson, "Vowel Amplitude variation during sentence production," in Proc. ICASSP'97, vol.3, pp.1627-1630, 1997.
5. J.C.Lee, S.H.Kim, Minsoo Hahn, "Intonation Processing for TTS Using Stylization and Neural Network Learning Method," in Proc. ICSLP'96, pp. 1377-1380, 1996.
6. H. Oh, Y.J. Lee, "A modified error function to improve the error back-propagation algorithm for multilayer perceptrons," ETRI J., vol.17, pp.11-22, 1995.