Pitch Contour Conversion Using Slanted Gaussian Normalization Based on Accentual Phrases*

Ki Young Lee · Myung Jin Bae**** · Ho Young Lee*** · Jong Kuk Kim****

ABSTRACT

This paper presents methods using Gaussian normalization for converting pitch contours based on prosodic phrases along with experimental tests on the Korean database of 16 declarative sentences and the first sentences of the story of "The Three Little Pigs". We propose a new conversion method using Gaussian normalization to the pitch deviation of pitch contour subtracted by partial declination lines: by using partial declination lines for each accentual phrase of pitch contour, we avoid the problem that a Gaussian normalization using average values and standard deviations of intonational phrase tends to lose individual local variability and thus cannot modify individual characteristics of pitch contour from a source speaker to a target speaker. From the results of the experiments, we show that this slanted Gaussian normalization using these declination lines subtracted from pitch contour of accentual phrases can modify pitch contour more accurately than other methods using Gaussian normalization.

Keywords: declination line, pitch deviation, prosodic phrase, accentual phrase

1. Introduction

Voice conversion requires transformation of all perceptually important aspect of the human voice: pitch, loudness, timbre and timing(tempo and rhythm). Tempo has more to do with overall speed while rhythm is more about local variations in speed. Timbre deals with how the voice itself sounds, while the other aspects reflect how a person speaks. There are many researchers who investigate how to convert pitch and timing. In practice, by varying pitch contours, a speaker who converses or reads can present not only state of emotion but also the meaning of sentences. A conversion of prosody features including pitch contour therefore plays an important role to express the desired characteristics of a

* This work was supported by the Korean Science and Engineering Foundation, grant no. R01-2002-000-00278-0.

** Dept. of Information Communication Engineering, Kwandong University
*** Dept. of Linguistics, Seoul National University
**** Dept. of Information Communication Engineering, Soongsil University

speaker and the meaning of an utterance[1, 2].

There are two approaches to pitch contour conversion. The one is a statistical approach such as Gaussian normalization, the other is a dynamic programming method using non-linear time warping based on pitch contours from a training sentence database[3, 4]. This dynamic programming method requires a large training database of utterances spoken by at least two speakers. The main idea of the statistical method for pitch contour is to convert the pitch frequency from a source to a target speaker using average values and standard deviation. We have already proposed a method to compensate local pitch variability of a target speaker using Gaussian normalization for each accentual phrase[5]. However, a Gaussian normalization using statistical average values and standard deviations tends to lose individual local variation and thus cannot modify individual characteristics of pitch contour from a source speaker to a target speaker[6].

This paper presents a new conversion method using slanted Gaussian normalization. The first processing of this method is to subtract the residues of pitch contour by a partial declination line of pitch contour for a source speaker. To compensate for local pitch variations, the second processing is to transform the residues by Gaussian normalization and should fit these transformed residues to the partial declination line of a target speaker based on a accentual phrases along with time scaling. Experiments are carried out for the 16 declarative sentences and the first sentences of the story "The Three Little Pigs" used for the fairy tale narration system[7]. These sentences are uttered by 5 males and 5 females for narrating a story to children. The experimental results show that the slanted Gaussian normalization proposed by this paper can modify pitch contour more accurately than the other methods using Gaussian normalization.

This paper consists of four major parts. In section 2, we overview and describe the declination line and the partial declination lines of pitch contour based on Korean prosodic phrases. The proposed pitch contour conversion methods are described in section 3. Experimental results are presented in section 4, and finally a conclusion is given in section 5.

2. Pitch model for Korean prosody

2.1 Korean prosody: intonational phrase and accentual phrase

The intonational phrases(IP) of Sun-Ah Jun is a prosodic unit which corresponds to the intonational phrase of Nespor and Vogel, and is characterized by an intonation contour made up of two tonal levels H(igh) and L(ow). The intonational contour of the IP is derived from two constituents: the pitch accent and the phrase tone. Thus, the phrase

accent marks the boundary of intermediate phrase which are smaller units than the IP. The smaller units than the IP are accentual phrase(AP) which are submit of the IP[8, 9].

The AP is marked by an F0 contour. Although the F0 contour has various patterns according to pragmatic meaning such as focus, topic, etc. and to dialects in Korean, Seoul dialect has the basic pattern of LH or LHLH according to the number of syllables which are contained in an AP, and if two H tones appear in an AP the second one is higher than the first one. As another characteristics, the AP's contours show the down-step or declination phenomenon in an IP. That is, the L tone of an AP is lower than that of the previous AP and the H tone is lower than that of the previous AP. The last AP's contour of an IP is overridden by the boundary tone of the IP. Using these characteristics, we can set up the basic pitch pattern of APs as follows:

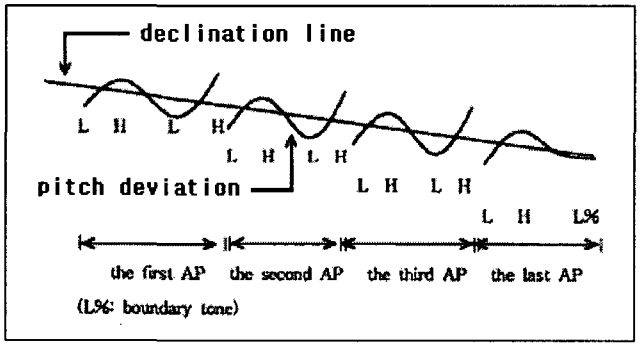

Figure 1. The basic pitch pattern of APs in an IP.

For Korean, the accentual phrase(AP) and the intonational phrase(IP) are linguistically significant. Experimental results indecate that Sun-Ah Jun's suggestion is valid in reading sentences [10].

2.2 Pitch model

As described in figure 1, the basic pitch contour of speech is sloped by the down-step effect. Futhermore, it is known that in a declarative phrase the pitch decreases overall. This is referred to as the pitch declination. Hence it is more accurate to determine declination lines of both speakers and to model the variations of the pitch around that line as normally distributed deviations.

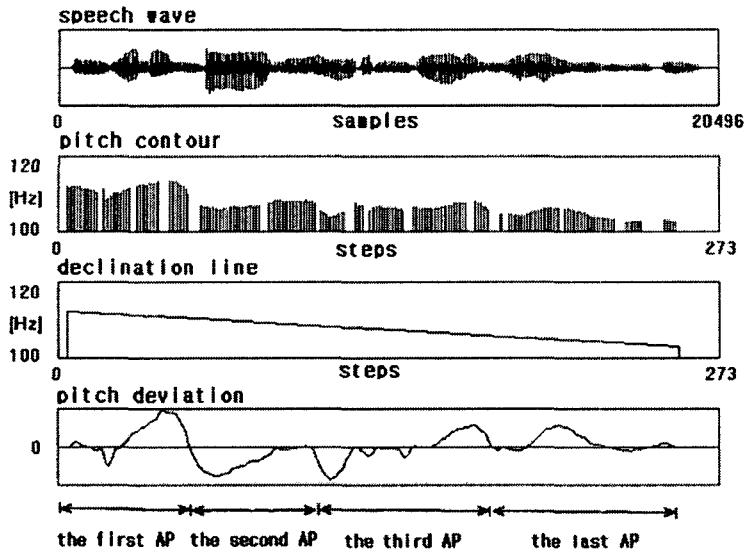

Figure 2. A simple pitch model: a declination line and pitch deviation.

We can consider a simple pitch model based on a declination line and pitch deviation such as LH, LHLH etc. according to the number of syllables for each AP as shown in figure 2. This declination line can be described simply as follows, where $p(t_0)$ is the pitch frequency of beginning time t_0, and $p(t_N)$ is the one of ending time t_N.

$$D(t) = p(t_0) + (t - t_0) \cdot \frac{p(t_N) - p(t_0)}{t_N - t_0} \tag{1}$$

$$\Delta p(t) = p(t) - D(t) \tag{2}$$

The pitch deviation can be modelled by subtracting a declination line from the pitch contour as described in the equation (2). This paper develops pitch contour conversion algorithms by the Gaussian normalization using these two models.

3. Gaussian normalizations for pitch contour conversion

The Gaussian normalization is a basic algorithm which is referred to as a Gaussian normalization in an IP in this paper. The presented approaches are divided by two levels of Korean prosodic phrases: intonational phrase and accentual phrase. The first presented normalization method is combined with a declination line of pitch contour in an IP,

referred to as declined Gaussian normalization in an IP. The second one is a Gaussian normalization within each AP to compensate for local pitch variations, referred to as accentual Gaussian normalization in each AP. The last one performs pitch contour conversion according to each AP by subtracting the declination line of the source speech, normalizing through a Gaussian algorithm and fitting converted pitch deviation to the declination line of the target speech, referred to as slanted Gaussian normalization in each AP.

3.1 Gaussian normalizations in an IP and each AP

The basic method of Gaussian normalization involves matching the average pitch and the standard deviation of pitch of a given source speaker to those of a target speaker for each IP. Assume that pitch measurement values are Gaussian random variables, where the average pitch and standard deviation of pitch of the source speaker before pitch conversion are μ^S and σ^S respectively, and the average pitch and the standard deviation of pitch of the target speaker are μ^T and σ^T respectively. Then given a pitch value of a source speaker, the modified pitch value $p^{S\to T}(t)$ is computed as

$$p^{S\to T}(t) = \frac{p^S(t) - \mu^S}{\sigma^S} \cdot \sigma^T + \mu^T \tag{3}$$

Accentual phrases(APs) are constituents of IP. There is a strong correlation between syntactic and prosodic phrases in Korean language. Korean syntactic phrases are divided in orthography by a space, and are in general in accordance with APs. Within an IP, an AP that is characterized by a pitch contour pattern LH (low-high) includes three syllables at maximum, and another AP that is characterized by a pitch contour pattern LHLH including four syllables at least. The last AP is a boundary tone that is different from the LH pattern.

The accentual Gaussian algorithm makes use of the local pitch patterns of the APs and carry out pitch contour conversion according to every AP by Gaussian normalization at a time. Then given a pitch value $p^{S_i}(t)$ in the i-th AP of a source speaker, the converted pitch value $p^{S_i\to T_i}(t)$ is computed using equation (3).

3.2 Declined and slanted Gaussian normalization

The declined Gaussian normalization algorithm begins with the assumption that the pitch contour of an IP has a down-step trend which can be fitted by a declination line. The algorithm therefore makes use of the declination line structure and applies Gaussian normalization only to the residual pitch values reduced from subtracting the declination

line in the pitch contour of each speaker's sentence. At last, the converted pitch residues of an IP for each speaker are fitted to the other's declination line.

Then the pitch residues or deviation $\Delta p(t)$ of each speaker are calculated as the equation (2). The pitch deviation or residues $\Delta p^{S}(t)$ and $\Delta p^{T}(t)$ of the source and the target speaker are modeled as two Gaussian random variables and Gaussian normalization is applied to obtain the converted residue $\Delta p^{S\to T}(t)$ by equation (3). Finally the modified pitch value is computed as

$$p^{S\to T}(t) = \Delta p^{S\to T}(t) + D^{T}(t) \tag{4}$$

We can consider about APs that the pitch values for each AP are converted in the same manner as the method which begins with the assumption that the pitch contour of each AP has a any sloped trend which can be fitted by each slanted line. Although this slanted line shouldn't have the declined trend such as the declination line, we assume that the slanted trend of each AP can be described by a declination line using equation (1). That is, slanted Gaussian normalization can perform pitch contour conversion according to every AP by subtracting slanted line $D^{S_i}(t)$ from $p^{S_i}(t)$ of the pitch contour of the i-th AP for a source speaker, normalizing through Gaussian algorithm by equation (3) and overlap-adding the converted pitch deviation $\Delta p^{S_i\to T_i}(t)$ to the slanted line $D^{T_i}(t)$ of target speech using equation (4) within every AP.

4. Experimental results and evaluation

Scripts used for data collection were composed of 16 sentences of declarative sentence and the first sentence of "The Three Little Pigs" story used for the fairy tale narration system[7]. Two male speakers of standard Korean read the declarative sentences in their natural style without any guideline. The first sentence of the story was read by speakers of 5 males and 5 females. Speech data were obtained by means of a 16-bit AD converter with an 11 kHz sampling rate.

4.1 Conversion results

Figure 3 shows the original speech wave and pitch contour of a source speaker and a target speaker which have three APs. The vertical lines on the pitch contour in figure 3 (b) and (d) show the boundary of each AP. The dotted lines and digits between the two boundaries represent the average pitch level and its value for each AP, respectively. The

pitch contour of the source speaker in figure 3(b) has the average pitch of 162[Hz] and the standard deviation of 9.33 and the higher pitch tone is revealed at the rightmost of the second AP. The pitch contour of a target speaker in figure 3(d) has the average pitch of 155[Hz] and the standard deviation of 12.54 and higher pitch tone reveals at the right part of the first AP. We can evaluate briefly the extent of pitch conversion by looking at the time varying shape of the pitch contour in figures 4 and 5.

Figure 4 shows the conversion results carried out by the Gaussian normalization. Figure 4(a) is the wave converted from the source speaker's speech wave to the target speaker's by Gaussian normalization in an IP. In figure 4(b), we can see that the converted pitch contour has the same values as the average and standard deviation of the target speaker's ones. However the higher tone is not revealed at the right part of the first AP like in figure 3(d). Figure 4(c) is the wave converted by a declined Gaussian normalization in an IP. And the average value and standard deviation of pitch contour in figure 4(d) are smaller than those in figure 4(b) because of subtracting the declination line. And the higher tone appear in the rightmost of the second AP. This means the trend of pitch contour doesn't appear when declined Gaussian normalization.

Figure 5 shows the conversion results carried out by normal and slanted Gaussian normalizations. Figure 5(a) is the wave converted from the source speaker's speech wave to the target speaker's by Gaussian normalization in each AP. In figure 5(b), we can see that the converted pitch contour has similar values to the average and standard deviation of a target speaker's ones. And the higher tone appears in the right part of the first AP like in figure 3(d). Figure 5(c) is the wave converted by slanted Gaussian normalization in each AP. And the average value and standard deviation of pitch contour in figure 5(d) is little different from those in figure 5(b) because of subtracting the each slanted line. And the higher tone appear in the right part of the first AP. This means the trend of pitch contour can appear more accurately by slanted Gaussian normalization in each AP than the other normalizations.

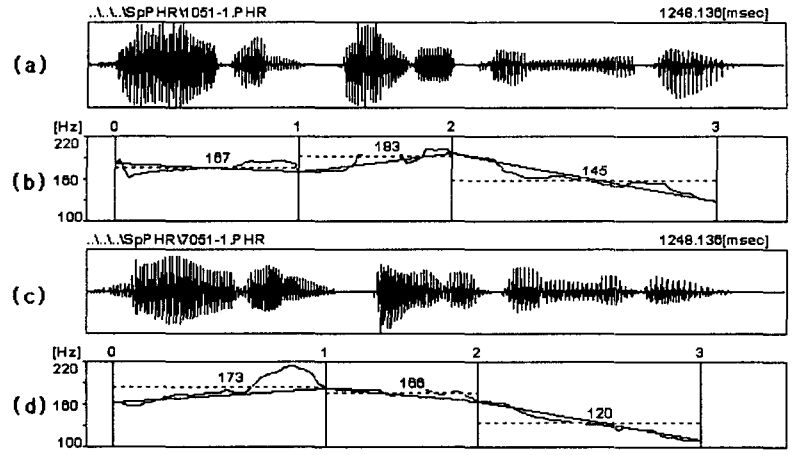

Figure 3. Original speech and pitch contour

(a) Wave of a source speaker (b) Pitch contour of (a)
(c) Wave of a target speaker (d) Pitch contour of (b)

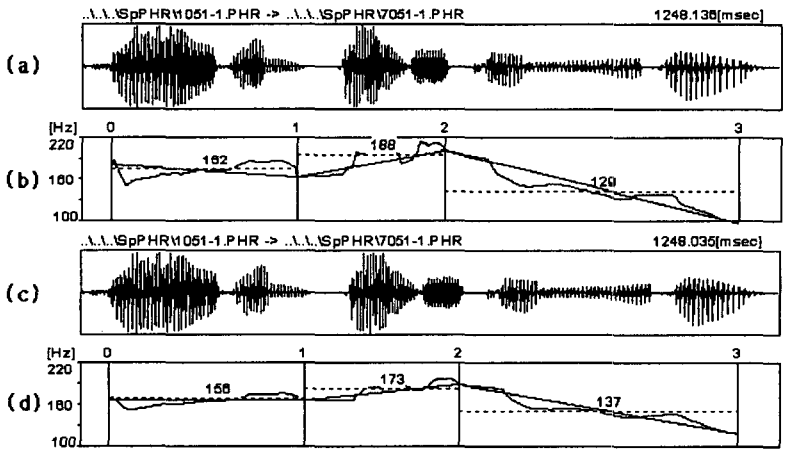

Figure 4. Speech and pitch contour converted in an IP

(a) Wave after Gaussian normalization (b) Pitch contour of (a)
(c) Wave after declined Gauss normalization (d) Pitch contour of (b)

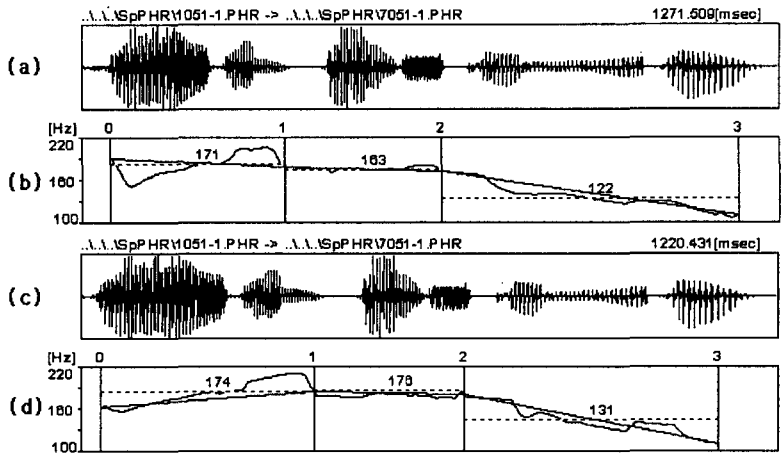

Figure 5. Speech and pitch contour converted in each AP

(a) Wave after accentual Gaussian normalization (b) Pitch contour of (a)
(c) Wave after slanted Gaussian normalization (d) Pitch contour of (b)

4.2 Evaluation

For evaluating the test, we made the tester's window as depicted in figure 6 to add the score 1 for the best similar converted speech of the four conversion methods. Conversion 1 means the Gaussian normalization in an IP. Conversion 2 means the declined Gaussian normalization in an IP. Conversion 3 means the accentual Gaussian normalization in each AP. Conversion 4 means the slanted Gaussian normalization in each AP.

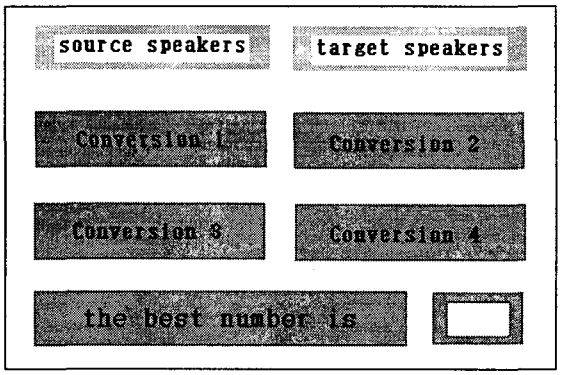

Figure 6. Tester's window for subject evaluation

In the tester's window, let the testers click the button to listen to several speech uttered by source speakers, target speakers and four conversion methods. After listening to several speeches through the window, the tester can write the number of the best

similar speech in the empty box of the bottom. Then the evaluation score can be increased by clicking the "the best number is" button. We use three utterances to evaluate converted speech subjectively. Testers are five males and females. Hence a perfect score is 30. Table 1 shows the evaluation scores of the testers.

Table 1. Evaluation scores

conversion methods	1	2	3	4
score	0	2	7	21
[%]	0	6.7	23.3	70.0

In the case of Gaussian normalization as the basic conversion method 1 in table 1, although the average pitch value and standard deviation are very similar to those of the target's, the score of subjective evaluation is the worst of the four conversion methods. the evaluation score of slanted Gaussian normalization in conversion method 4 is higher than the others.

5. Conclusion

We presented three methods for pitch contour conversion application with a basic method of Gaussian normalization. These methods are mainly based on a statistical approach using the average values and standard deviation of pitch contour in two prosodic phrases: intonational phrase(IP) and accentual phrase(AP). And we use a simple model of pitch contour as a declination line and pitch deviation.

In the level of IP, the results of the basic algorithm of Gaussian normalization and the other algorithm using a declination line of pitch contour show that it is not good to modify pitch contour to the target speaker, since the IP unit is too long to compensate pitch variation in one sentence. This includes several APs that have multiple patterns of tonal levels. In the level of AP, the results of the accentual and the slanted Gaussian normalization show that they are promised for converting pitch contour to a target speaker, since the unit of AP is adjustable to compensate for the basic pattern or pitch deviation.

Experimental results show that the proposed algorithm of slanted Gaussian normalization at the level of APs is capable of modifying pitch contours more accurately than the algorithms of the others, since within each AP the ranges of pitch variation is much less than the range of pitch variation in the IP and the slant trend can be compensated by fitting pitch deviation to the declination line of each AP.

References

[1] Akagi, M. & Ienaga, T., 1995. "Speaker Individualities in Fundamental Frequency Contours and Its Control", Proc. EuroSpeech '95, 439-442.

[2] Kuwabara, H. & Sagisaka, Y., 1995. "Acoustic Characteristics of Speaker Individuality : Control and Conversion", Speech Communication, Vol. 16, 165-173.

[3] Arslan, L. M. & Talkin, D., 1998. "Speaker Transformation using Sentence HMM based Alignments and Detailed Prosody Modification", Proc. ICASSP '98, Vol. 1, 289-292.

[4] Chappel, D. T. & Hansen, J. H. L.. 1998. "Speaker-Specific Pitch Contour Modeling and Modification", Proc. ICASSP '98, Vol. 1, 885-888.

[5] Lee, K. Y. & Zhao, Y., 2004. "Statistical conversion algorithms of pitch contours based on prosodic phrases", Proc. Speech Prosody 2004.

[6] van Santen, J. P. H., 1997. "Prosodic Modeling in Text-to-Speech Synthesis", Proc. EuroSpeech '97, KN 19-KN 28.

[7] Park, Won & Bae, Myung Jin. 2000. "A study on the fairy tale narration system with key-word exchange", Proceeding of the 2000, Korean Signal Processing Conference, Vol. 13, No. 1, 819-822.

[8] Nespor, M. & Vogel, I., Prosodic Phonology, Dordrecht : Foris Publication.

[9] Jun, Sun-Ah. 1993. The Phonetics and Phonology of Korean Prosody, Ph. D. Dissertation, The Ohio State University.

[10] Lee, K. Y. & Song, M. S., 1999. "Automatic Detection of Korean Accentual Phrase Boundaries", The Journal of Acoustic Society of Korea, Vol. 18, No.1E, 27-31.

Received: February 10, 2004
Accepted: March 10, 2004

▲ Ki Young Lee
Dept. of Information Communication Engineering, Kwandong University
Imcheonli San 7 Yangyang Eub Yangyang Kun, Kangwon Do, 215-800, Korea
Tel: +82-33-670-3413
E-mail: kylee@kwandong.ac.kr

▲ Ho Young Lee
Dept. of Linguistics, Seoul National University
Sinlim 9 Dong San 56-1, Kwanak Ku, Seoul, 151-742, Korea
Tel: +82-2-880-6166
E-mail: hylee@snu.ac.kr

▲ Myung Jin Bae
Dept. of Information Communication Engineering, Soongsil University
Sangdo 5 Dong 1-1, Dongjak Ku, Seoul, 156-743, Korea
Tel: +82-2-820-0902
E-mail: mjbae@ssu.ac.kr

▲ Jong Kuk Kim
Dept. of Information Communication Engineering, Soongsil University
Sangdo 5 Dong 1-1, Dongjak Ku, Seoul, 156-743, Korea
Tel: +82-2-824-0906
E-mail: kokjk@hanmail.net