Statistical Conversion Algorithms of Pitch Contours Based on Prosodic Phrases

*Ki Young Lee, Yunxin Zhao**

Dept. of Information Communication Engineering, Kwandong University, S. Korea
*Dept. of Computer Science, University of Missouri-Columbia, USA
leeki@missouri.edu *zhaoy@missouri.edu

Abstract

Pitch contour of a speech utterance plays an important role in expressing speaker's individuality and meaning of the utterance. In performing speech conversion from a source speaker to a target speaker, it is important that the pitch contour of the source speaker's utterance be converted into that of the target speaker. This paper investigates statistical algorithms of pitch contour conversion for Korean language. The algorithms are based on Gaussian normalization, and its combination with a declination-line modeling of pitch contour. Pitch contour conversion are investigated at two levels of prosodic phrases: intonation phrase and accentual phrase. Experimental results show that the algorithm of Gaussian normalization within accentual phrases is significantly more accurate than the algorithms for intonational phrases in pitch contour conversion.

1. Introduction

Speech conversion is a process to transform an utterance spoken by a source speaker in such a way that it is perceived to be spoken by a target speaker. Through varying pitch contours, a speaker who converses or reads can present not only state of emotion but also meaning of sentence. A conversion of prosody features including pitch contour therefore plays an important role to express desired characteristics of a speaker and meaning of an utterance.

Pitch contour has been used to make high quality synthetic speech through TTS (text-to-speech) systems that are capable of expressing speaker individuality. Psychoacoustic experiments support the theory that pitch contours contain speaker individuality [1, 2]. In TTS, intonation as expressed by pitch contours is generated in accordance with the unit of sentence or other structures defined by such systems [3, 4], and prosodic phrases have been shown beneficial to naturalness of synthetic speech [5].

Currently there are two approaches to pitch contour conversion. One is a statistical method using Gaussian normalization, the other is a dynamic programming method using non-linear time warping based on pitch contours from a training sentence database [6,7]. The statistical method is simple because the average pitch value of a given speaker can be mapped to that of a target speaker by a Gaussian normalization. However the method is insufficient to capture local pitch variations as perceived in the utterance of the target speaker. The dynamic programming method requires a large database of utterances spoken by at least two speakers.

In this paper, we propose new methods of pitch contour conversion based on prosodic phrases where the statistical method of Gaussian normalization is improved to compensate for local pitch variations. In the first method, Gaussian normalization is performed on pitch contour of each intonational phrase based on the 1^{st} and 2^{nd} order pitch statistics. In the second method, the pitch contour of an intonation phrase is first fitted by a declination line and the resulting pitch residues are then compensated for by Gaussian normalization. The third method performs Gaussian normalization on pitch contour of each accentual phrase based on the 1^{st} and 2^{nd} order pitch statistics. The latter method is able to accurately convert pitch contour of a source speaker to pitch contour of a target speaker that is rich of local variation structure.

The rest of the paper is organized as the following. The prosody property of Korean language is overviewed in section 2. The proposed pitch contour conversion methods are described in section 3. Experimental results are presented in section 4, and a conclusion is made in section 5.

2. Prosodic phrases

Nespor and Vogel [8] proposed that human languages have a universal hierarchical structure that consists of seven prosodic units, including syllables, feet, phonological words, clitic group, phonologicaphrases, intonational phrases and phonological utterance. These units are closely related to the prosodic and phonological rules appropriate to each language. Sun-Ah Jun [9] proposed that not all seven prosodic units of Nespor and Vogel are necessary for each language, but to each language there are a few units that are linguistically significant. For Korean, she suggested, accentual and intonational phrases are linguistically significant. Experimental results support her suggestion to be valid in read sentences [10]. In this paper, we develop methods of pitch contour conversion based on accentual and intonational phrases.

3. Pitch contour conversion methods

The proposed methods of converting the pitch contours of a given speaker to those of a target speaker are summarized in Table 1. The first two algorithms perform pitch contour conversion in the unit of intonational phrases (IP). One algorithm is Gaussian normalization, and the other is a combination of declination line fitting followed by Gaussian

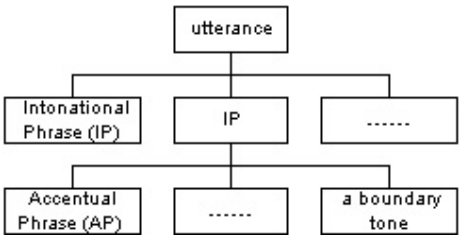

Figure 1: *Korean prosodic phrases*

normalization, referred to as declined Gaussian. The last algorithm performs pitch contour conversion according to accentual phrases by Gaussian normalization, referred to as accentual Gaussian. The PSOLA [11] technique is used for synthesizing speech waveform after the described pitch contour conversion.

Table 1: *Pitch contour conversion algorithms*

Prosodic phrase	Algorithm	Approach
IP	Gaussian	Gaussian normalization
	Declined Gaussian	Declination line fitting and Gaussian normalization
AP	Accentual Gaussian	Gaussian normalization in accentual phrase

3.1 Pitch contour conversion with IP

3.1.1 Gaussian normalization algorithm

The method of Gaussian normalization involves matching the average pitch and the standard deviation of pitch of a given source speaker to those of target speaker for each intonational phrase. Assume that pitch measurement values are i.i.d. 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 p_t^S of a source speaker, the modified pitch value $p_t^{S->T}$ is computed as

$$p_t^{S->T} = \frac{p_t^S - \mu^S}{\sigma^S} \cdot \sigma^T + \mu^T \quad (1)$$

In implementing this algorithm, pitch tracking is first performed on training sentences from both the source and target speakers, and estimation is then made on the mean and standard deviation of pitch values of each intonational phrase for each speaker. It is easily shown that the converted pitch values $p_t^{S->T}$ by equation 1 has the mean and standard deviation matched to those of the target speaker.

3.1.2 Declined Gaussian algorithm

The algorithm of Gaussian normalization has limited capability in pitch contour conversion due to the underlying assumption that pitch values are i.i.d. Gaussian random variables within each intonatioanl phrase. The declined Gaussian normalization algorithm begins with the assumption that the pitch contour of each intonational phrase has a declination trend which can be fitted by a line. The algorithm therefore makes use of the declination line structure and applies Gaussian normalization only to the residue pitch values resulting from subtracting the declination line in the pitch contour of the target speaker from the pitch contours of the source speaker and the target speaker, respectively. In this formulation, the pitch contour of an intonational phrase of the target speaker is first fitted by the declination line

$$D_t^T = p_{t_0}^T + (t - t_0)\frac{p_{t_N}^T - p_{t_0}^T}{t_N - t_0}$$

where $p_{t_0}^T$ and $p_{t_N}^T$ are the pitch values at a starting time t_0 and an ending time t_N, respectively for the target speaker. Then the pitch residues Δp_t^S and Δp_t^T of the source and target speakers are computed as $\Delta p_t^S = p_t^S - D_t^T$ and $\Delta p_t^T = p_t^T - D_t^T$. The residues Δp_t^S and Δp_t^T are modeled as two i.i.d. Gaussian random variables and Gaussian normalization is applied to obtain the converted residue $\Delta p_t^{S->T}$ by equation 1. Finally the modified pitch value is computed as

$$p_t^{S->T} = \Delta p_t^{S->T} + \left\{ p_{t_0}^T + (t - t_0)\frac{p_{t_N}^T - p_{t_0}^T}{t_N - t_0} \right\} \quad (2)$$

3.2 Accentual Gaussian algorithm

Accentual phrases are constituents of intonational phrase. In Korean, syntactic phrases are divided in orthography by a space, and are in general in accordance with accentual phrases. There is a strong correlation between syntactic and prosodic phrases.

Within an intonational phrase, an accentual phrase that is characterized by a pitch contour pattern LH (low-high) includes three syllables at maximum, and an accentual phrase that is characterized by a pitch contour pattern LHLH includes four syllables at least. The last accentual phrase is a boundary tone that is different from the LH pattern.

The accentual Gaussian algorithm makes use of the local pitch patterns of the accentual phrases and carry out pitch conversion on one accentual phrase at a time.

4. Experimental results and evaluation

Speech data were obtained at 10 kHz sampling rate. Script used for data collection was composed of 16 sentences with all sentences declarative. There were 4 sentences with 4 accentual phrases each, 4 sentences with 5 accentual phrases each, and 8 sentences of 3

accentual phrases each. The total number of accentual phrases added up to 60. Two male speakers of standard Korean read the script in their natural style without any guideline. Prosodic phrase boundaries were hand marked at the levels of intonational phrases and accentual phrases.

4.1 Conversion results

Figure 2 shows the conversion results on a pair of intonatioanl phrases by the three algorithms of Table 1. PSOLA algorithm was used to re-synthesize speech samples according to the converted pitch contours. Figure 2 (a) and (b) are speech waveform and pitch contour of a source speaker, and (c) and (d) are the speech waveform and pitch contour of a target speaker. The vertical lines in (b), (d), (f), (h) and (j) are hand-marked boundaries of accentual phrases.

The speech waveform and pitch contour after Gaussian normalization are shown in (e) and (f). The speech after Gaussian normalization has the same average pitch and standard deviation of pitch as those of the target speaker in each intonatioanl phrase. However, the resulting pitch contours of accentual phrases are very different from the target ones.

Figure 2 (g) and (h) are speech waveform and pitch contour modified by using the declined Gaussian algorithm. The result shows that the starting and ending pitch values of the modified speech are identical to those of the target speakers, but the algorithm failed to capture large local variations of pitch contours.

The results from using accentual Gaussian algorithm are shown in Figure 2 (i) and (j). It is observed that this algorithm is able to accurately modify pitch contours even for large local pitch variations.

4.2 Evaluation

Both subjective and objective measures may be used to evaluate the results of pitch contour conversion. In subjective evaluation, human subjects would listen to pitch-modified speech data and their opinions are collected for scoring each method. In objective evaluation, pitch contour error in the modified speech data relative to that of the target speaker is directly measured. Since in certain cases the scale of pitch contour modification are not large enough to be clearly perceived in listening tests, the objective measure was used to quantify the error of pitch conversion. Define the pitch error in the *i*-th accentual phrase as

$$e_i = \frac{1}{M}\sum_{m=0}^{M}\left(\left|\mu_i^T - \mu_i^{S_m->T}\right|\right) \qquad (3)$$

where $\mu_i^{S_m->T}$ represents the average pitch of i-th accentual phrase for the modified speech from a m-th source speaker to a target speaker and μ_i^T means the average pitch of i-th accentual phrase for the target speech, respectively, with $1 \le i \le N$ and N is the number of accentual phrases. Table 2 shows error comparison for all 16 sentences with accentual phrases. In the case of Gaussian normalization, the average error is about 5.8 Hz. In the declined Gaussian algorithm, the average error is about 15.7 Hz. In the accentual Gaussian algorithm, the error is converged to 0.00 Hz, since Gaussian normalization was applied to each accentual phrase..

Table 2: *Comparison of pitch contour errors in Hz for the accentual phrases***.**

i-th AP	1st	2nd	3rd	4th	5th	Avg
Gaussian	5.1	8.5	10.0	3.6	1.9	5.8
Declined Gaussian	14.4	12.9	19.1	14.5	17.5	15.7
Accentual Gaussian	0.3	0.5	0.4	0.2	0.4	0.4
Average	6.6	7.3	9.8	6.1	6.6	

5. Conclusion

The same sentence spoken by two speakers in general has different prosodic characteristics including duration, intensity and tone. In the current work, statistical algorithms of pitch contour conversion are proposed to modify the pitch contours of prosodic phrases from a source speaker to those of a target speaker. Experimental results showed that the proposed algorithm of Gaussian normalization at the level of accentual phrases is capable of modifying pitch contours more accurately than the algorithms for intonational phrases, since within each accentual phrase the ranges of pitch variation is much less than the range of pitch variation in the intonational phrase.

Acknowledgement

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

References

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

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

[3] A. Kain; M.W. Macon, 1998. Spectral Voice Conversion for Text-To-Speech Synthesis. *Proc. ICASSP'98, Vol. 1*, 285-288.

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

[5] Y. J. Kim; H. J. Byeon; Y. H. Oh, 1999. Prosodic Phrasing in Korean; Determine Governor, and then Split or Not. *Proc. EuroSpeech'99*, 539-542.

[6] L. M. Arslan; D. Talkin, 1998. Speaker

Transformation using Sentence HMM based Alignments and Detailed Prosody Modification. *Proc. ICASSP'98, Vol. 1*, 289-292.

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

[8] M. Nespor; I. Vogel, 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] K. Y. Lee; M. S. Song, 1999. Automatic Detection of Korean Accentual Phrase Boundaries. *The Journal of Acoustic Society of Korea, Vol. 18, No.1E*, 27-31.

[11] E. Moulines; F. Charpentier, 1990. Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones". *Speech Communication Vol. 9*, 453-467.

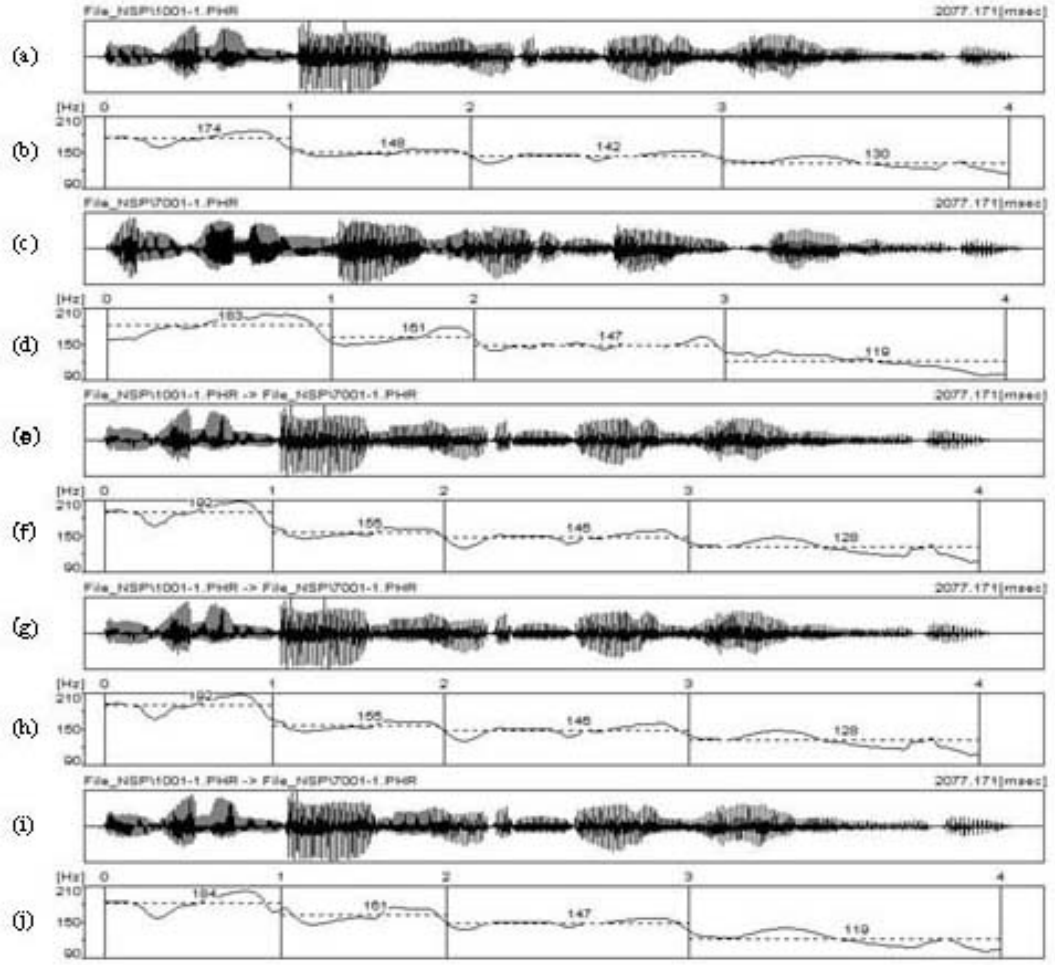

Figure 2. *Results of pitch contour conversion*

(a) Speech waveform of a source speaker
(b) Pitch contour of (a)
(c) Speech waveform of a target speaker
(d) Pitch contour of (c)
(e) Speech waveform after Gaussian normalization
(f) Pitch contour of (e)
(g) Speech waveform after Declined Gaussian normalization
(h) Pitch contour of (g)
(i) Speech waveform after Accentual Gaussian normalization
(j) Pitch contour of (i)