THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.23, NO.1E 2004, 3 pp. 10~15

Statistical Approaches to Convert Pitch Contour Based on Korean Prosodic Phrases

Ki Young Lee
Dept. of Information Communication Engineering, Kwandong University
(Received January 15 2004; accepted March 13 2004)

Abstract

In performing speech conversion from a source speaker to a target speaker, it is important that the pitch contour of the source speakers utterance be converted into that of the target speaker, because pitch contour of a speech utterance plays an important role in expressing speaker's individuality and meaning of the utterance. This paper describes statistical algorithms of pitch contour conversion for Korean language. Pitch contour conversions are investigated at two l evels of prosodic phrases: intonational phrase and accentual phrase. The basic algorithm is a Gaussian normalization [7] in intonational phrase. The first presented algorithm is combined with a declination-line of pitch contour in an intonational phrase. The second one is Gaussian normalization within accentual phrases to compensate for local pitch variations. Experimental results show that the algorithm of Gaussian normalization within accentual phrases is significantly more accurate than the other two algorithms in intonational phrase.

Keywords: *Speech conversion, Pitch contour, Goussian normalization, Intonation phrases, Accentual phrases*

I. 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. Psychoacoustic experiments support the theory that pitch contours contain speaker individuality[1, 2].

Pitch contour has been used to make high quality synthetic speech through TTS (text-to-speech) systems that are capable of expressing speaker individuality, and intonation as expressed by pitch contours is generated in accordance with the unit of sentence or other structures defined by such systems[3, 4]. In TTS systems, 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 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[6, 7]. The statistical method of a Gaussian normalization is easy to process because the average pitch value of a given speaker can be mapped to that of a target speaker. However this method is insufficient to capture local pitch variations as perceived in the utterance of the target speaker. The dynamic programming method requires a large training database of utterances spoken by at least two speakers.

The purpose of this study is to present two algorithms

Corresponding author: Ki Young Lee (kylee@kd.ac.kr)
Dept. of Information Communication Engineering, Kwandong University, 7 san Imcheon-ri Yangyang-Eup Yangyang-Kun Kangwou-Do 215-802 korea

based on prosodic phrases for converting the pitch contour of a sentence for the sake of imparting perceptually important characteristics of a desired speaker, where the statistical method of Gaussian normalization is improved to compensate for local pitch variations. The basic algorithm is a Gaussian normalization that is performed on pitch contour using the average pitch and the standard deviation of pitch statistics. In the first presented algorithm, the pitch contour of an intonation phrase is first fitted by a declination line and the resulting pitch residues are then converted by Gaussian normalization. The second one performs Gaussian normalization on pitch contour of every accentual phrase for each sentence to compensate for local pitch variation. Experiments are carried out for several declarative sentences uttered by a source and a target speaker, and pitch contour error within every accentual phrase of modified speech relative to that of the target speaker is measured to evaluate their converting abilities because the scale of pitch contour modification are not large enough to be clearly perceived in listening tests. The result shows that the second 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.

II. Prosodic Phrases of Korean

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, phonological phrases, 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.

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 intonationl 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. The pitch accent is a pitch event phonologically linked to a particular stressed syllable in an utterance. The phrase tone is an autosegment which exists independently of lexical entries, and consists of phrase accents and a boundary tone. The phrase accent occur after the rightmost pitch accent, and a boundary tone occurs at the right edge and optionally at the left edge of the IP. 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. In sum, the natural utterance is composed of the hierarchical structure which has APs and IPs as its constituents.

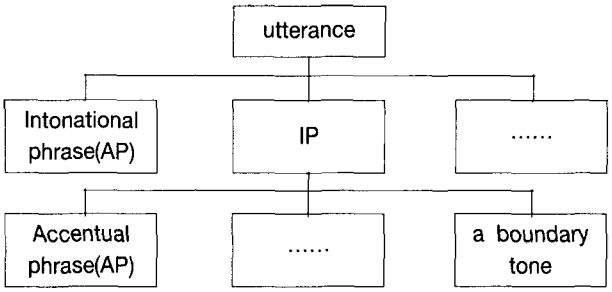


Figure 1. Korean prosodic phrases.

For Korean, accentual phrases (APs) and intonational phrases (IPs) are linguistically significant. Experimental results support her suggestion to be valid in reading sentences[10]. This paper develops the statistical algorithms of pitch contour conversion based on prosodic phrases of Korean.

III. Statistical Algorithms of Pitch Contour Conversion

The presented 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 prosodic unit of IP. The Gaussian normalization is a basic algorithm, and the other one is a combination of declination line fitting followed by Gaussian normalization, referred to as declined

Table 1. Pitch contour conversion algorithms.

Prosodic phrase	Algorithm	Approach	Assumption	Main Idea
Intonational Phrase	Gaussian	statistics	Gaussian distributuion	Gaussian normalzaion
	Declined Gaussian	declination line, statistics	pitch declination base line	effective distribution
Accentual Phrase	Accentual Gaussian	accentual phrase, statistics	accentual phrase	effective distribution, prosody phrase

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

3.1. Pitch Contour Conversion in 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 IP. 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 of a source speaker, the modified pitch value $p_t^{S\to T}$ is computed as

$$p_t^{S\to T}=\frac{p_t^S-\mu^S}{\sigma^S}\cdot\sigma^T+\mu^T \tag{1}$$

In implementation this algorithm, pitch tracking is first performed on training sentences from both the source and target speaker, and estimation is then made on the mean and standard deviation of pitch values of each IP for each speaker. It is not complicated to show that the converted pitch values $p_t^{S\to T}$ by equation (1) has the mean and standard deviation matched to those of the target speaker in IP.

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 IP. 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 each speaker's sentence. In this formulation, the pitch contour of an IP of each speaker is first fitted by the declination line,

$$D_t=p_{t_0}+(t-t_0)\cdot\frac{p_{t_N}-p_{t_0}}{t_N-t_0} \tag{2}$$

where p_{t_0} and p_{t_N} are the pitch values at a starting time, t_0 and an ending time, t_N, respectively.

Then the pitch residues $\triangle p_t$ of each speaker are calculated as $\triangle p_t=p_t-D_t$. The residues $\triangle p_t^S$ and $\triangle p_t^T$ of the source and the target speaker are modeled as two i.i.d. Gaussian random variables and Gaussian normalization is applied to obtain the converted residue $\triangle p_t^{S\to T}$ by equation (1). Finally the modified pitch value is computed as

$$p_t^{S\to T}=\triangle p_t^{S\to T}+\left\{p_{t_0}^T+(t-t_0)\frac{p_{t_N}^T-p_{t_0}^T}{t_N-t_0}\right\} \tag{3}$$

3.2. Accentual Gaussian Algorithm in AP

Accentual phrases are constituents of IP. In Korean, syntactic phrases are divided in orthography by a space,

and are in general in accordance with APs. There is a strong correlation between syntactic and prosodic phrases in Korean language. 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 includes four syllables at least. The last AP is a boundary tone that is different from the LH pattern[9, 10].

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 $p_i^{S_i}$ in the i-th AP of a source speaker, the modified pitch value $p^{S_i \to T_i}$ is computed as

$$p^{S_i \to T_i} = \frac{p_i^{S_i} - \mu^{S_i}}{\sigma^{S_i}} \cdot \sigma^{T_i} + \mu^{T_i} \quad (4)$$

where μ^{S_i}, σ^{S_i} and μ^{T_i}, σ^{T_i} are the average pitch and the standard deviation of pitch of the source speaker and target speaker according to the i-th AP, respectively.

IV. 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 declarative sentences. 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 IPs and APs.

4.1. Conversion Results

Figure 2 shows the conversion results performed by the three algorithms of Table 1. Speech waveform is shown in figure (a), and figure (b) is the pitch contour of a source speaker, A, and (c) and (d) are the speech waveform and pitch contour of a target speaker, B. The vertical lines in (b), (d), (f), (h) and (j) are hand-marked boundaries of prosodic phrases such as IP and APs, where the boundaries of IP are the same of one spoken sentence and the smaller units than the IP are APs.

The speech waveform and pitch contour after Gaussian normalization are shown in figure (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 IP. Figure (g) and (h) are speech waveform and pitch contour modified by using the declined Gaussian algorithm. The result shows that only the starting and ending pitch values of the modified speech are identical to those of the target speakers. However, in the view of the AP unit, the resulting pitch contours are different from the target ones.

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.

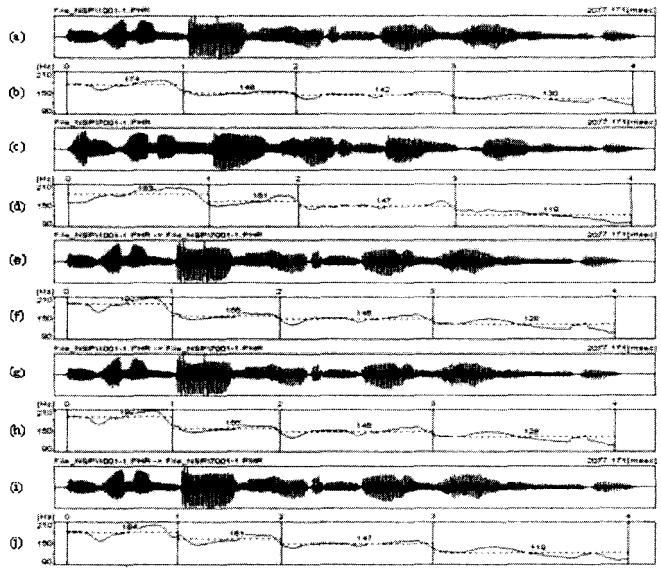


Figure 8. Results of pitch contour conversion.
(a) Speech waveform of a source speaker
(b) Pitch contour of (a)
(c) Speech waveform of a target speaker
(d) Pitch contour of (c)
(e) Speech waveform after Gaussian normalization
(f) Pitch contour of (e)
(g) Speech waveform after Declined Gaussian normalization
(h) Pitch contour of (g)
(i) Speech waveform after Accentual Gaussian normalization
(j) Pitch contour of (i)

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.

4.2.1. Objective Evaluation

Since in certain cases the scale of pitch contour modification are not large enough to be clearly perceived in listening tests, the objective measure is used to quantify the error of pitch conversion. Define the pitch error in the i-th accentual phrase as

$$e_{\mu}^{i}=\frac{1}{M}\sum_{m=0}^{M}(|\mu_{m}^{T_i}-\mu_{m}^{S_i\rightarrow T_i}|) \tag{5}$$

$$e_{\sigma}^{i}=\frac{1}{M}\sum_{m=0}^{M}(|\sigma_{m}^{T_i}-\sigma_{m}^{S_i\rightarrow T_i}|) \tag{6}$$

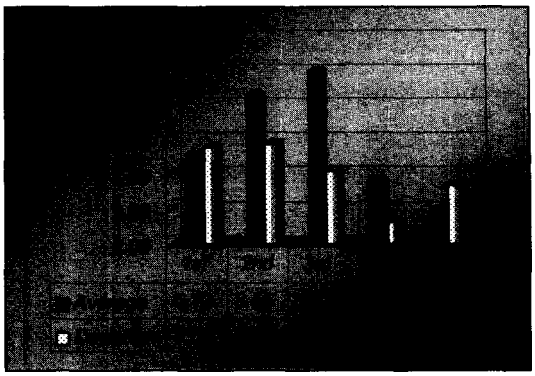
where $\mu_{m}^{S_i\rightarrow T_i}$ and $\sigma_{m}^{S_i\rightarrow T_i}$ represents the average pitch and the standard deviation of pitch for i-th accentual phrase of the modified speech from a source speaker to a target speaker, respectively, with $1\leq m\leq M$ and M is the number of the spoken texts.

Figure 3 shows comparisons of pitch error values computed by equation (5) and (6) performed by three algorithms. In figure 4, the comparison of three algorithms is presented by their average of pitch error values for all APs.

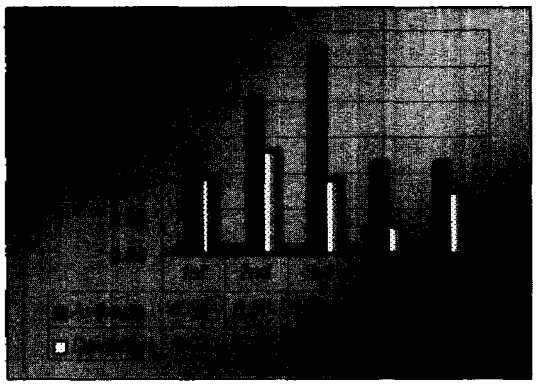
In the case of Gaussian normalization, the average of pitch error for all accentual phrases is about 5.82 and the deviation error is about 4.05. In the declined Gaussian algorithm, the average of pitch error is about 7.14 and the deviation is about 3.67. In the accentual Gaussian algorithm, the errors are converged near to 0, because this algorithm uses a conversion unit as an AP smaller than an IP. Since within each AP the ranges of pitch variation is much less than the range of pitch variation in the IP, this proposed algorithm using APs can modify pitch contours more accurately than others using IPs.

4.2.2. Subjective Evaluation

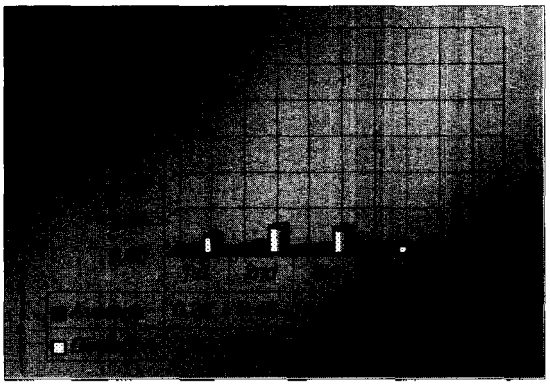
To subjectively evaluate conversion performance, one forced-choice(ABX) test was carried out. In this test material, we took 4 sentences in the experimental script, and each sentence is consisted of 4 APs.



(a) Gaussian Normalization



(b) Declined Gaussian algorithm



(c) Accentual Gaussian algorithm

Figure 3. Pitch error comparisons of each algorithm

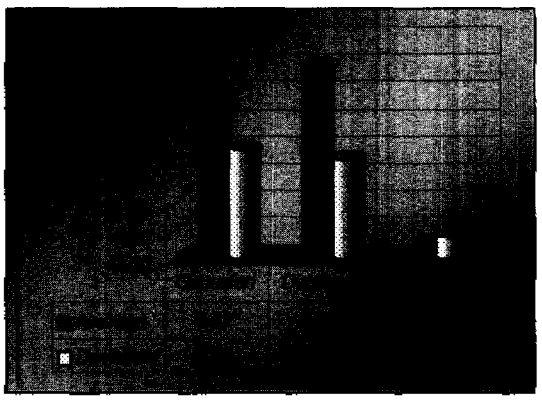


Figure 4. Average error comparison of 3 algorithms.

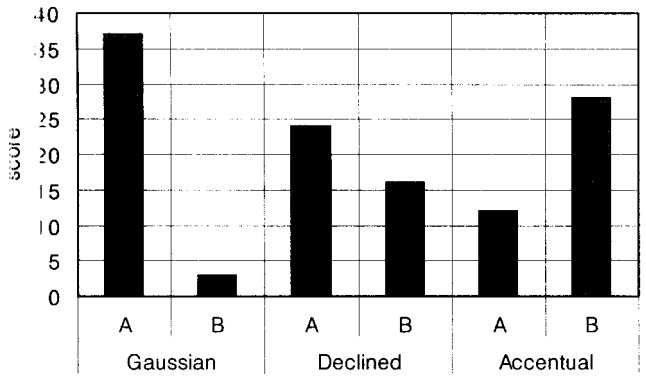


Figure 5. Results of ABX test.

In the ABX experiment, let each listener hear several speeches of source, target and converted speech, and A and B were speech utterances spoken by a source speaker and a target spaeker. X was the converted speech from the tone of A to the one of B by each algorithm. When listener listen to three Xs by 3 algorithm, we ask, "is the tone of X closer to the one of A or to the one of B?" For example, if he selected A, the score is increased by 1. Since we used 4 sentences to ABX test for 10 listeners, the highest score is 40 (= 4 sentences * 10 listeners). The result of the tone perceptual tests were shown in figure 5. In this experiment, 10 listeners were not easy to decide who was speaking, but they could recognise that the tone of X by the accentual Gaussian algorithm would be different from A and more similar to B than those converted by the other algorithms.

V 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.

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 a target speaker, since the IP unit is too long to compensate pitch variation in one sentence including several APs that have multiple patterns of tonal levels.

Experimental results show that the proposed algorithm of Gaussian normalization at the level of APs is capable of modifying pitch contours more accurately than the algorithms for IPs, since within each AP the ranges of pitch variation is much less than the range of pitch variation in the IP.

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,"Speaker Individualities in Fundamental Frequency Contours and Its Control", Proc. EuroSpeech'95, pp. 439-442, Sep. 1995
2. H. Kuwabara, Y. Sagisaka,"Acoustic Characteristics of Speaker Individuality: Control and Conversion", Speech Communication, **16**, pp.165-173, 1995
3. A. Kain, M.W. Macon,"Spectral Voice Conversion for Text-To-Speech Synthesis", Proc. ICASSP'98, **1**, pp. 285-288, 1998
4. J. P. H. van Santen,"Prosodic Modeling in Text-to-Speech Synthesis", Proc. EuroSpeech'97, KN 19-KN 28, 1997
5. Y. J. Kim, H. J. Byeon, Y. H. Oh,"Prosodic Phrasing in Korean: Determine Governor, and then Split or Not", Proc. EuroSpeech'99, pp.539-542, 1999
6. L. M. Arslan, D. Talkin,"Speaker Transformation using Sentence HMM based Alignments and Detailed Prosody Modification", Proc. ICASSP'98, **1**, pp. 289-292, 1998
7. D. T. Chappel, J. H. L. Hansen,"Speaker-Specific Pitch Contour Modeling and Modification", Proc. ICASSP'98, **1**, pp. 885- 888, 1998
8. M. Nespor, I. Vogel, Prosodic Phonology, Dordrecht : Foris Publication
9. Jun, Sun-Ah, The Phonetics and Phonology of Korean Prosody, Ph. D. Dissertation, The Ohio State University, 1993
10. K. Y. Lee, M. S. Song, "Automatic Detection of Korean Accentual Phrase Boundaries", The Journal of Acoustic Society of Korea, **18**(1E), pp.27-31, 1999
11. E. Moulines, F. Charpentier,"Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", Speech Communication 9(5,6) pp.453-467, 1990

[Profile]

• Ki Young Lee
The journal of the acoustic society of Korea Vol.22, No.3, 2003.