말소리와 음성과학 제1권 제4호 (2009)
pp. 47~59

Building a Sentential Model for Automatic Prosody Evaluation

Yoon, Kyuchul[1]

ABSTRACT

The purpose of this paper is to propose an automatic evaluation technique for the prosodic aspect of an English sentence uttered by Korean speakers learning English. The underlying hypothesis is that the consistency of the manual prosody scoring is reflected in an imaginary space of prosody evaluation model constructed out of the three physical properties of the prosody considered in this paper, namely: the fundamental frequency (F0) contour, the intensity contour, and the segmental durations. The evaluation proceeds first by building a prosody evaluation model for the sentence. For the creation of the model, utterances from native speakers of English and Korean learners for the target sentence are manually scored by either native teachers of English or Korean phoneticians in terms of their prosody. Multiple native utterances from the manual scoring are selected as the "model" native utterances against which all the other Korean learners' utterances as well as the model utterances themselves can be semi-automatically evaluated by comparison in terms of the three prosodic aspects [7]. Each learner utterance, when compared to the multiple model native utterances, produces multiple coordinates in a three-dimensional space of prosody evaluation, each axis of which corresponds to the three prosodic aspects. The 3D coordinates from all the comparisons form a prosody evaluation model for the particular sentence and the associated manual scores can display regions of particular scores. The model can then be used as a predictive model against which other Korean utterances of the target sentence can be evaluated. The model from a Korean phonetician appears to support the hypothesis.

Keywords: automatic prosody evaluation, prosody evaluation model, English, Korean learners, Praat

1. Introduction

The evaluation of the knowledge of a foreign language can be performed in various ways. One way that has been frequently employed by teachers of English education involves asking questions with multiple choices. However, this type of questions should be inappropriate in testing the prosodic aspects of an uttered sentence. The question that one might ask as to the prosodic evaluation of an uttered sentence is how do we test the prosodic properties.

One way would be to evaluate the prosodic aspects of an utterance using the knowledge of English grammar. The target utterance could be analyzed in terms of the knowledge and graded accordingly. This knowledge-based approach should work fine as long as there is consensus among the human evaluators as to the knowledge of the grammar involved. Even if there were consensus, it would be inevitable that the evaluators be different in their opinions in the actual application of the agreed grammatical knowledge. For example, how prominent should a stressed syllable be to be safely called a stressed syllable? Things will get worse if it is the F0 contour of an utterance.

In order to overcome issues like the inter-evaluator reliability, researchers have worked to develop various measures that reflect the prosodic aspects of utterances. Speech corpora can be used to develop these measures and the established measures can be used to assess the pronunciation proficiency of unknown utterances. One such measure has to do with rhythm metrics ([4] among others), which are based on segmental durations. Measures such as this have the advantage that once established, they can be applied consistently to a large amount of utterances for the purpose of evaluating the rhythmic property of the involved

[1] Yeungnam University kyoon@ynu.ac.kr

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2007-327-A00310).

Received: October 7, 2009
Revision received: December 10, 2009
Accepted: December 10, 2009

utterances.

However, this approach is likely to suffer from the same kind of problem as that encountered in the knowledge-based approach because the measures such as the rhythm metrics are "abstracted" concepts. For example, the F0 contour, the intensity contour and the segmental durations can be regarded as raw, less abstracted measures. These are physical measures that linearly involve human psychology. On the contrary, stress is an abstracted concept. How each of the raw physical properties interact to form what we perceive as stress in a word or a sentence can be an issue of intense debate. By its nature, abstracted concept needs to be tested and verified to become established as a reliable psychological measure.

The two issues of knowledge and abstraction need to be resolved if a reliable prosody evaluation technique is to be established. An alternative would be to find a different source of knowledge and to avoid abstraction. Rather than trying to establish the knowledge that can be applied to evaluate an utterance, one can borrow the whole utterance produced by a native speaker of English and simply use the knowledge implicit in the utterance. If one native speaker's "model" utterance is insufficient, we can borrow the same sentence uttered by different native speakers, collecting more "model" utterances. Then the evaluation process will have to involve some kind of comparison between the model native utterances and the learner utterance produced by a Korean speaker. How to compare the two types of utterances is another issue that will be discussed shortly. The evaluation by comparison, if proven to be effective, can be a good way to bypass the problem of the knowledge-based approach.

An alternative to the issue of abstraction would be to adopt the three physical properties of the prosody and use them in the comparison process of the prosody evaluation. The viability of this approach has been explored by Yoon [7]. His experiment involved utterances whose prosody, in the form of the three physical properties, had been artificially distorted. Although his experiment showed promising results, this non-abstraction approach needs to be verified with more natural utterances.

The problem of how to compare the two types of utterances, i.e. between the native "model" utterances and the learner utterance, can be simplified if the number and durations of the segments involved between the two types are rendered identical. If the dimension of the segmental durations are fixed using the PSOLA algorithm [3], the other two dimensions, i.e. the F0 and intensity contours, can be easily compared (See <Figure 1>). For example, a sample sentence uttered by a native speaker of English and a Korean student can be very different in terms of the three prosodic properties. Once the number and durations of the component phones are made identical, the comparison in terms of the F0 and intensity contours between the two utterances becomes a matter of simple comparison between the matching pitch and intensity points [7]. The comparison in the segmental durations does not need any durational modifications.

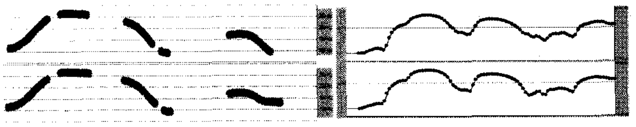
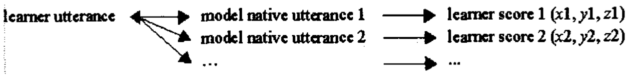


Figure 1. The comparison of the pitch points (left panel) and the intensity points (right panel) between the original utterance (upper contours) and its distorted version (lower contours) [7].

1.1. Prosody evaluation by comparison with multiple model utterances

The use of multiple model utterances from native speakers of English is one of the key ideas of this paper. If a learner utterance is scored prosodically by being compared to each of the model native utterances, it means that a learner utterance will have as many learner scores as the number of model utterances. The same procedure can be applied to the model native utterances. Each model utterance can be scored prosodically by comparison with the other model utterances, also producing multiple model scores. The actual scoring of the learner utterance will take place by comparing these sets of multiple scores from the learners and the native speakers of the model utterances.



The multiple scores per utterance approach appears to be plausible because there is no fixed way of uttering a particular sentence. Even if a sentence is uttered in a neutral setting without focusing or emphasizing any part of it, there can be many slightly different but still correct ways of uttering the sentence. Therefore it is natural that there should be a range of prosodic realizations which can be judged as correct by native speakers of English.

1.2. Building a 3-dimensional sentential model for prosody evaluation

Since we will be considering the three physical properties of prosody, each of the multiple scores, whether they are from learners' or model native utterances, will also be composed of

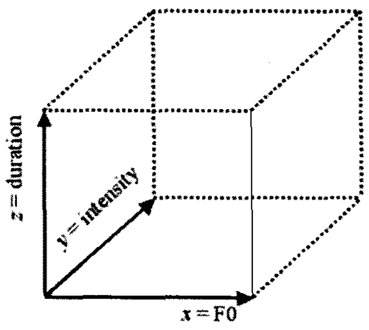


Figure 2. The imaginary 3-dimensional space of prosody evaluation.

three values; the first one from the F0 contour comparison (x-coordinate), the second one from the intensity comparison (y-coordinate) and the third one from the duration comparison (z-coordinate). The three-valued score will occupy a single point in a three-dimensional space of prosody evaluation (See <Figure 2>). The x-coordinate will represent the comparison in the F0 contour axis, the y-coordinate in the intensity contour axis, and the z-coordinate in the segmental durational axis. The three-valued prosody scoring system based on comparisons between the model native utterances and the learner utterances can be visualized as an imaginary three-dimensional space as in <Figure 2>.

Comparisons among the model native utterances will yield a range of score points whose x-, y-, and z-coordinate values will not differ very much compared with the values from Korean learners. Thus the score points will be placed near the axis origin of the three axes, i.e. the black[2] points near the lower left front corner of the prosody evaluation space in <Figure 3>. Assuming that the learner utterances are not as good as the model native utterances, comparisons between the two types of utterances can be expected to yield score points farther away from the axis origin, i.e. the red points toward the upper right back corner of the space. The more the prosodic discrepancy between the two types, the farther away the score points will be placed from the axis origin. The two possible scenarios can be seen in <Figure 3>.

If the localization of the score points is as evident as in <Figure 3>, we could see this space as a prosody evaluation model for that particular sentence and use it to evaluate unknown utterances of the same sentence. For example, if the manual prosody scores for the red points in <Figure 3> are known to be 1 out of a five-point scale, 1 beiluate uworst and 5 beiluate ubestody scan unknown utterancee manuascore points from the automatic comparison1 out multiple utterances of theiout multiplred points, the utteranceecould be gtern the score 1 acc for g to the imagr tareprosody evaluation model. In an id ml case, the point groups for each of the manual scores will occupy their own domain in the 3D model, makiluate upredictterap uttof the evaluation model straluut. If the domains of the manual scores overlap too much, the predictive power of the model will be reduced.

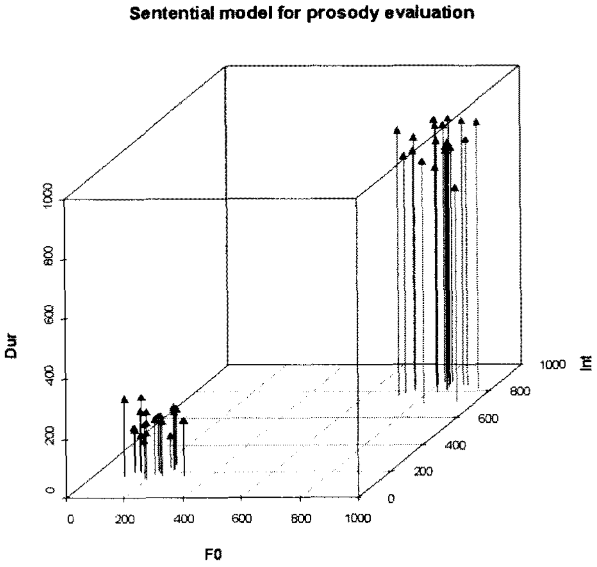


Figure 3. The two possible scenarios of the score points in the prosody evaluation. The arrowheads represent the actual score points. Vertical lines are drawn to help identify the locations of the points in the 3-dimensional space.

1.3. Interaction with segment evaluation

The evaluation model that we intend to build is for the prosodic aspects of an utterance. The prosodic aspects of an utterance, however, cannot be separated from its segmental aspects. It would be pointless and unrealistic, if not impossible, to judge the prosodic aspects of an utterance without considering any component segments. Even a trained phonetician would have difficulty performing such task. Therefore it is necessary that we control the segmental proficiency of an utterance in the evaluation. The prosody evaluation model could be built on utterances whose segment evaluation scores are the same. Then we will get multiple prosody evaluation models for a particular sentence depending on the manual scores of the segments.

2. Methods

2.1. Manual evaluation of sentential prosody

Two sets of utterances were used for the experiments. The sentences and their recordings are from [5].

[2] The color version of this article can be downloaded at the journal website.

Set A: "Miss Henry drank a cup of coffee."
Set B: "The dancing queen likes only the apple pies."

For the experiment involving Set A, the recordings were made from 296 speakers with the sampling frequency of 16kHz in a sound-proof studio([5]). The overall proficiency of all the utterances were evaluated by five native teachers of English, who did not participate in the recording. The evaluation was based on a five-point scale, 1 being the worst and 5 being the best. Note that the evaluation for Set A sentences was for the overall proficiency of the utterances involved, meaning that the segmental aspects as well as the prosodic aspects of an utterance were considered. Thus there was no control for the evaluation of the segmental aspects of an utterance. Therefore the experiment involving the Set A utterances aims to see if there is any correlation between the overall manual scoring and the 3D prosody evaluation model. Then the recordings were classified into four groups according to the evaluation scores. 10 utterances were selected from each of the following groups, 40 utterances in total. <Table 1> shows the actual scores of the forty utterances.

N5 : 10 native speakers of English
who were given the overall score point 5.
K5 : 10 Korean learners
who were given the overall score point 5 or almost 5.
K3 : 10 Korean learners
who were given the overall score point 3 or almost 3.
K1 : 10 Korean learners
who were given the overall score point 1 or almost 1.

For the other experiment involving Set B, the same recording and evaluating procedures were performed except that it was a Korean phonetician who did the evaluation. This time, the segmental aspects as well as the overall proficiency of each utterance were evaluated. A total of 30 utterances were selected from each of the following groups (Note that the group names for Set B are primed). <Table 2> shows the actual scores of the thirty utterances. The segment evaluation was controlled so that the Korean learners were all given the score point 4 for the segmental aspects.

N5' : 12 native speakers of English
who were given the overall score point 5.
K5' : 8 Korean learners
who were given the overall score points 5 or 4.
K2' : 12 Korean learners
who were given the overall score points 3 or 2.

Table 1. The overall scores of the 40 utterances for Set A sentence "Miss Henry drank a cup of coffee". Eval# (where # are numbers from 1 to 5) represents the native teacher evaluators who performed the evaluation.

Group	Spkr.	Eval1	Eval2	Eval3	Eval4	Eval5
N5	1	5	5	5	5	5
	2	5	5	5	5	5
	3	5	5	5	5	5
	4	5	5	5	5	5
	5	5	5	5	5	5
	6	5	5	5	5	5
	7	5	5	5	5	5
	8	5	5	5	5	5
	9	5	5	5	5	5
	10	5	5	5	5	5
K5	11	5	4	5	5	4
	12	5	5	5	5	4
	13	4	5	5	5	4
	14	5	5	5	5	5
	15	4	5	4	5	5
	16	4	5	5	5	4
	17	4	5	5	5	5
	18	4	5	4	5	5
	19	4	5	5	5	5
	20	4	5	5	4	5
K3	21	3	2	3	3	3
	22	3	3	3	3	4
	23	3	3	3	4	3
	24	3	3	3	3	2
	25	3	2	3	3	3
	26	3	3	4	3	3
	27	3	2	3	3	3
	28	3	3	3	3	2
	29	3	2	3	3	3
	30	3	3	3	4	3
K1	31	1	2	1	1	2
	32	1	1	2	2	1
	33	1	1	2	1	2
	34	1	1	2	1	1
	35	1	1	1	1	1
	36	1	1	1	1	1
	37	1	1	1	1	2
	38	1	1	1	1	1
	39	1	1	1	1	1
	40	1	1	1	2	1

Table 2. The segment and overall scores of the 30 utterances for Set B sentence "The dancing queen likes only the apple pies". The evaluation was performed by a Korean phonetician. Note that the segment evaluation scores were controlled among the Korean learner groups K5' and K2'.

Group	Spkr	Segment	Overall	Group	Spkr	Segment	Overall
N5'	1	5	5	K2'	21	4	3
	2	5	5		22	4	3
	3	5	5		23	4	3
	4	5	5		24	4	3

	5	5	5		**25**	4	3
	6	5	5		**26**	4	3
	7	5	5		**27**	4	3
	8	5	5		**28**	4	3
	9	5	5		**29**	4	2
	10	5	5		**30**	4	2
	11	5	5				
	12	5	5				
K5'	**13**	4	5				
	14	4	4				
	15	4	4				
	16	4	4				
	17	4	4				
	18	4	4				
	19	4	4				
	20	4	4				

The working hypothesis for the two experiments is that the consistency of the manual scoring is reflected in an imaginary space of prosody evaluation model constructed out of the three physical properties of the prosody; the F0 contour, the intensity contour, and the segmental durations. Since there was no control for the segment evaluation in the Set A experiment, a weaker correlation is expected than the Set B experiment which has the control.

The correlation can be regarded as each group having its own particular region in the 3D prosody evaluation space as shown in <Figure 3>. If the regions of each of the groups in the 3D space can be identified with a sufficient resolution, then the 3D model can be used to evaluate other utterances of the same sentence. If the 3D model from the Set B experiment had a sufficient resolution, it could be regarded as a viable prosody evaluation model for that particular sentence and used to score the prosodic aspects of other utterances of the same sentence. For example, if an unknown utterance version of the sentence belongs to a particular region in the 3D model, the utterance can be given the same prosody score point as the other utterances in the same region.

2.2. Semi-automatic evaluation of sentential prosody

This part of the experiment was where the semi-automatic evaluation of the prosodic aspects of an utterance took place. The segment labeling process was automatized by using a monophone HMM-based speech recognition engine[3].

Once the component segments of a learner utterance and the multiple model native utterances were identified and labeled, an additional segment adjustment step followed. The adjustment was necessary because the prosody evaluation technique introduced in [7] demands that the matching component segments be the same in their number. All these steps were performed in Praat [1].

Following the adjustment, three Praat scripts were executed onto a learner utterance along with its model native utterance. The comparisons by the three scripts yielded three coordinate values each of which was plotted on each of the three axes of the 3D prosody evaluation model (See <Figure 2>). As mentioned earlier, the comparisons were made with the other learner - model native utterance pairs, yielding multiple prosody score points for a particular learner.

2.3. A sentential prosody evaluation model based on evaluation by native teachers of English

Table 3. A sample coordinates of the overall proficiency score points for K5U1 utterance. N5 and K5 represent the group names and U# represents the utterance/speaker number.

	Difference in			
Comparison between	**F0 contour**	**Intensity contour**	**Segmental durations**	**Coordinates (x,y,z)**
N5.U1 - K5U1	899	142	408	(899, 142, 408)
N5.U2 - K5U1	360	92	190	(360, 92, 190)
N5.U3 - K5U1	377	159	183	(377, 159, 183)
N5.U4 - K5U1	206	81	151	(206, 81, 151)
N5.U5 - K5U1	291	153	826	(291, 153, 826)
N5.U6 - K5U1	251	113	563	(251, 113, 563)
N5.U7 - K5U1	346	120	532	(346, 120, 532)
N5.U8 - K5U1	299	114	343	(299, 114, 343)
N5.U9 - K5U1	282	92	216	(282, 92, 216)
N5.U10 - K5U1	716	178	183	(716, 178, 183)

The Set A utterances were used to build a sentential prosody evaluation model. Since there was no control on the segment evaluation part of the overall proficiency scores, even the best evaluation model cannot be regarded as reflecting only the prosodic aspects of the utterances involved.

In order to plot the overall proficiency score points of all the utterances in terms of the three prosodic properties, each of the 10 utterances from the four groups, i.e. N5, K5, K3 and K1, was semi-automatically evaluated by comparing them against each of the 10 utterances in the N5 group which was the select model native utterances. A batch Praat script produced 400 score points, all of which were plotted in a 3D space. A sample score points are shown in <Table 3>. For the calculation of the difference values, the method proposed in [7] was followed.

2.4. A sentential prosody evaluation model based on evaluation by a Korean phonetician

The Set B utterances were used to build another sentential

[3] The author would like to thank KT R&D Center for their help.

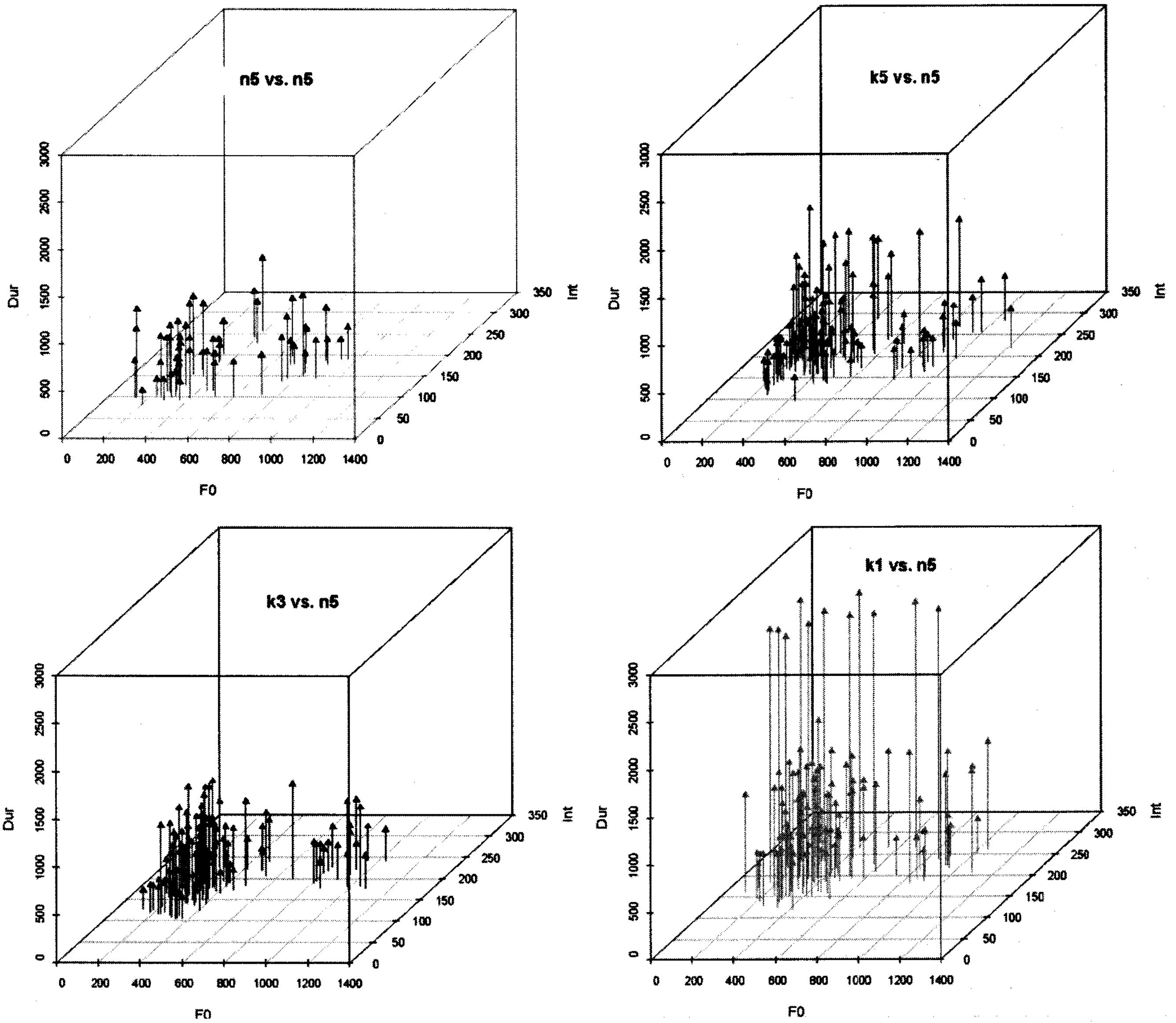


Figure 4. The 3D sentential prosody evaluation models based on evaluation by native teachers of English. The arrowheads represent the actual location of the score points. Vertical lines are drawn to help identify the locations of the points in the 3-dimensional space.

prosody evaluation model. As mentioned earlier, the utterances in this set were controlled, i.e. with the score of 4, in terms of the segment proficiency evaluation (Refer to <Table 2>). Thus, it is possible to assume that the overall scores are directly related to the prosodic proficiency evaluation.

Each of the utterances from the three groups, i.e. N5', K5' and K2', was semi-automatically evaluated with a batch Praat script by comparing them against the model native utterances, which was the N5' group in this case. This produced 360 score points, i.e. (12 + 8 + 10) x 12 = 360, and they were plotted in another 3D space.

2.5. A discriminant analysis of the prosody evaluation model

In order to verify the usefulness of the prosody evaluation model that was built from a Korean phonetician, a discriminant analysis was performed implemented in Praat. All the automatic prosody score points from the learners were divided into two groups, one for the training of the discriminant function and the other for the testing of the function. The training group contained all the automatic prosody scores and their manual scores from the two groups, i.e. K5' and K2'. The testing group contained only the two sets of automatic prosody scores from two learners, one of which was from K5' and the other from K2'. The discriminant function was derived from the training group. The manual prosody scores were predicted for the two learners of the testing

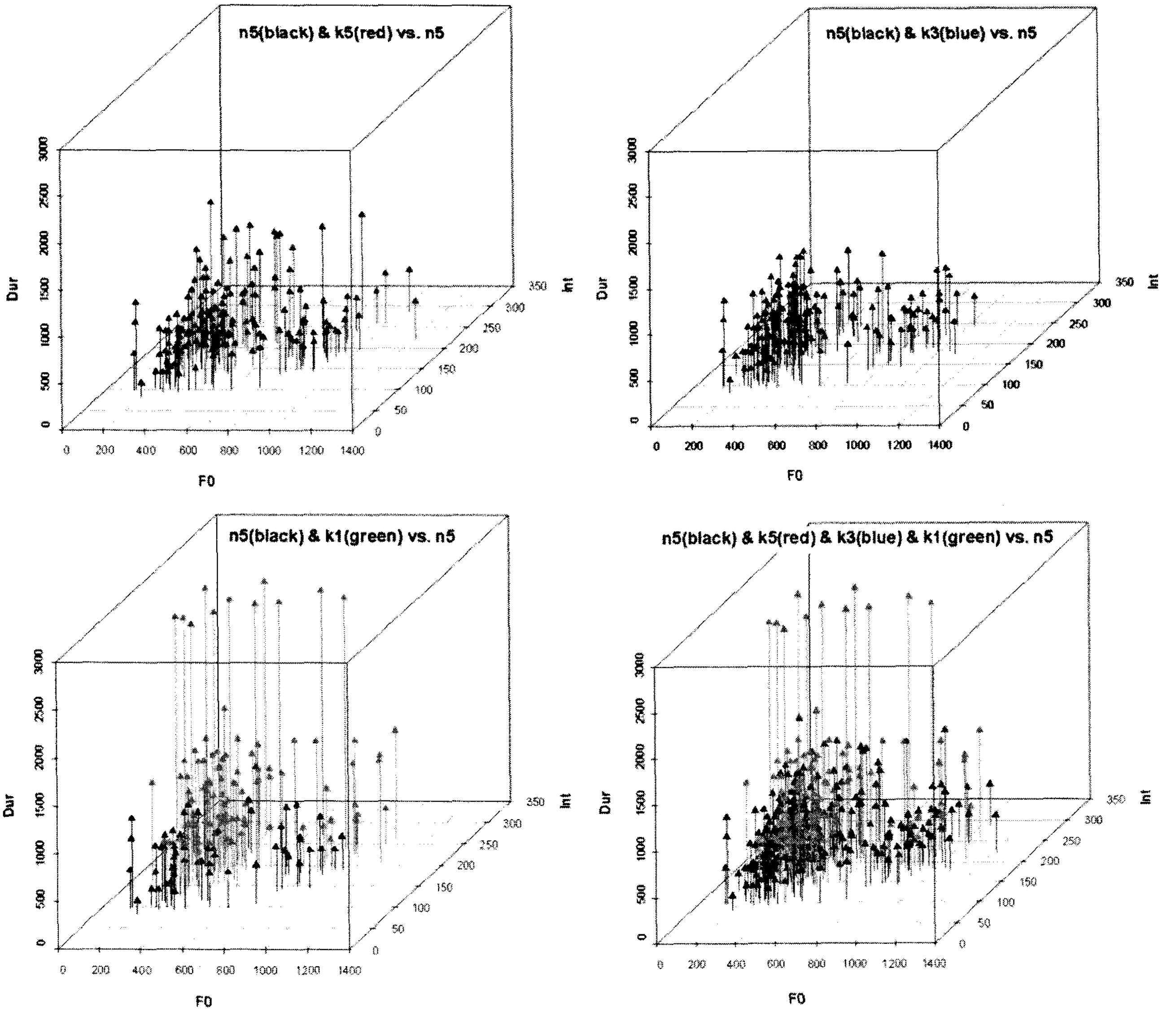


Figure 5. The 3D sentential prosody evaluation models based on evaluation by native teachers of English. The points from other groups are compared with the points from the N5 group. The points from the groups N5, K5, K3 and K1 are represented in black (all panels), red (upper-left), blue (upper-right) and green (lower-left) arrows respectively. All the groups are mixed in the lower-right plot.

group and a confusion matrix was produced. In addition, in order to examine the contribution of the three prosodic factors to the prosody evaluation, a multiple regression analysis and a principal component analysis(PCA) were performed.

3. Results

3.1. A 3D sentential prosody evaluation model by native teachers of English

The sentential prosody evaluation models based on the manual evaluation by native teachers of English and the semi-automatic evaluation by Praat scripts are given in <Figure 4>. The colors, i.e. black, red, blue, green, in each of the 3D models represent the manual scores and the arrowheads in each model represent the three-valued score points for that particular score group from the semi-automatic comparisons. The four prosody evaluation models in the form of a 3D scatterplot are for the Set A sentence "Miss Henry drank a cup of coffee". Each of the four models represents the overall proficiency scores of the N5(black), K5(red), K3(blue) and K1(green) groups and no control for the segmental proficiency evaluation was made.

The score points in each 3D scatterplot represents each of the scores points given manually by the native teachers of English who participated in the evaluation. For example, the points in the 3D scatterplot in the upper-left panel represent the comparisons among the model native utterances themselves that were given the score of 5 by other native teachers of English. The points in the 3D scatterplot in the upper-right panel represent the comparisons

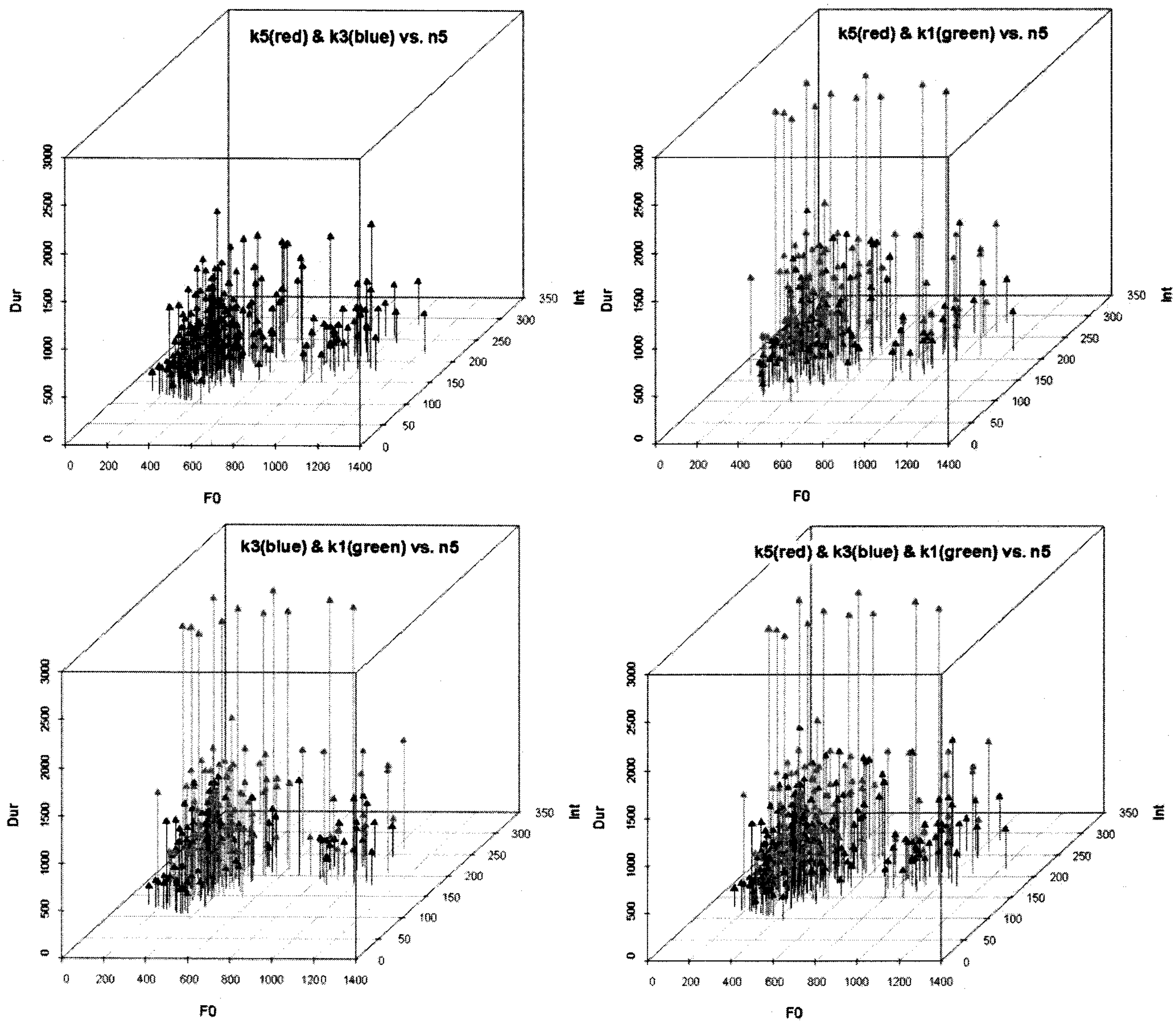


Figure 6. The 3D sentential prosody evaluation models based on evaluation by native teachers of English. The points from the learner groups K5, K3 and K1 are compared with each other. The upper-left panel is for K5(red) and K3(blue) points. The upper-right panel is for K5(red) and K1(green) points. The lower-left panel is for K3(blue) and K1(green) points. The lower-right panel is for all the points in the three groups.

between the model native utterances and the learner utterances with the score of (almost) 5. The points in the 3D scatterplot in the lower-left panel represent comparisons between the model native utterances and the other learner utterances with the score of (almost) 3 and those in the lower-right panel the score of (almost) 1. The relative position of the score points in each of the 3D plot represents the 3D coordinates from the semi-automatic prosody evaluation. All the scales are the same for the four plots.

The immediate difference between the N5 group and the rest of the plots for the K5, K3 and K1 groups, is that the points in N5 are relatively closer to the axis origin than the points in the other three plots. Such tendency is greater for the points in K5 and K3 groups than those in K1. Thus, N5, K5 and K3 seem to show a similar r ttern. It is understandable that some of the points in K1 are fe oher towe d the opposite of the axis origin because these are the ones with the worst score. Howeverter fovariation show a simpoints from the three learner models seems so great that it is he d to draw any conclusion from the plots. Notiw a sat there was no control for the segmental proficiency evaluation, it is not surprising that we get these confusing patterns.

This is more obvious if we examine the patterns in different combinations of the four groups as in <Figure 5>. The point groups in the upper-left and upper-right plots (for K5 and K3 points against N5 points) show similar patterns with many points clustered around each other. We can notice a slight degree of

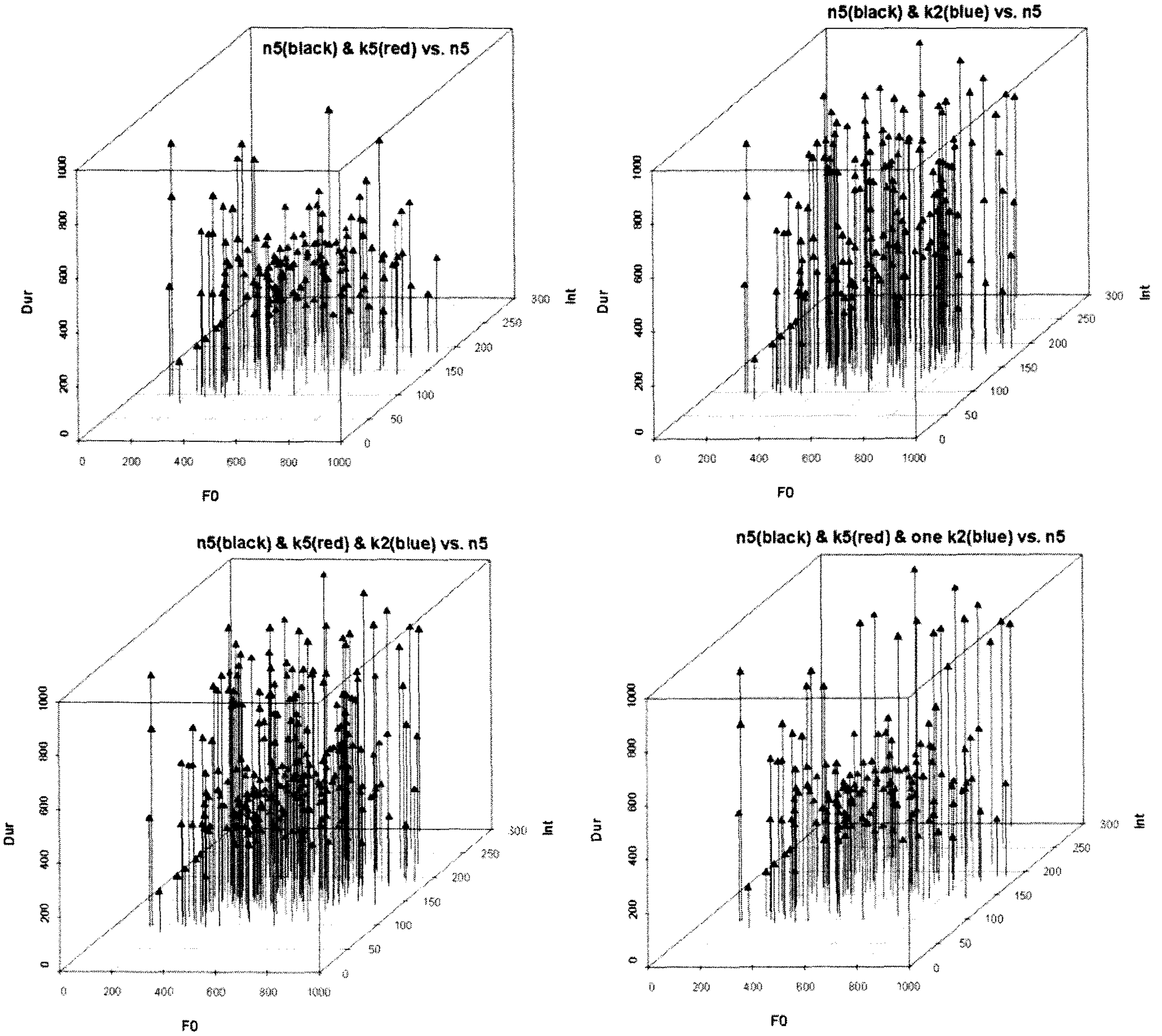


Figure 7. The 3D sentential prosody evaluation models based on evaluation by a Korean phonetician. The points from the model native group N5' (black) and the learner groups K5' (red) and K2' (blue) are compared with each other. The comparisons are between N5' and K5' (upper-left), N5' and K2' (upper-right), N5', K5' and K2' (lower-left) and N5', K5' and a single K2' point (lower-right).

separation between the two groups of points in the lower-left plot for K1 points against N5 points.

The comparison among the learner utterances gives a clearer picture (See <Figure 6>). The separation between the K5 points (red arrowheads in the upper-left plot) and the K3 points (blue arrowheads in the same plot) is not clear-cut. The separation between the K1 points (green arrowheads in the upper-right panel) and the K5 points (red arrowheads in the same panel) seems better than in the upper-left plot. Same tendency can be observed in the lower-left plot. Therefore, it can be said that even without any control for the segmental proficiency evaluation the learners' utterances in K1 can be somewhat differentiated from either those in K5 or in K3. However, the degree of separation does not seem to be good enough to be used as a predictive model for automatic prosody scoring.

3.2. A 3D sentential prosody evaluation model by a Korean phonetician

The sentential prosody evaluation models based on the manual evaluation by a Korean phonetician and the semi-automatic evaluation by Praat scripts are given in <Figure 7>. These prosody evaluation models are for the Set B sentence "The dancing queen likes only the apple pies". Each of the models can be said to represent only the prosody proficiency scores of the N5', K5' and K2' groups because the segmental proficiency evaluation scores were controlled. The segmental proficiency scores

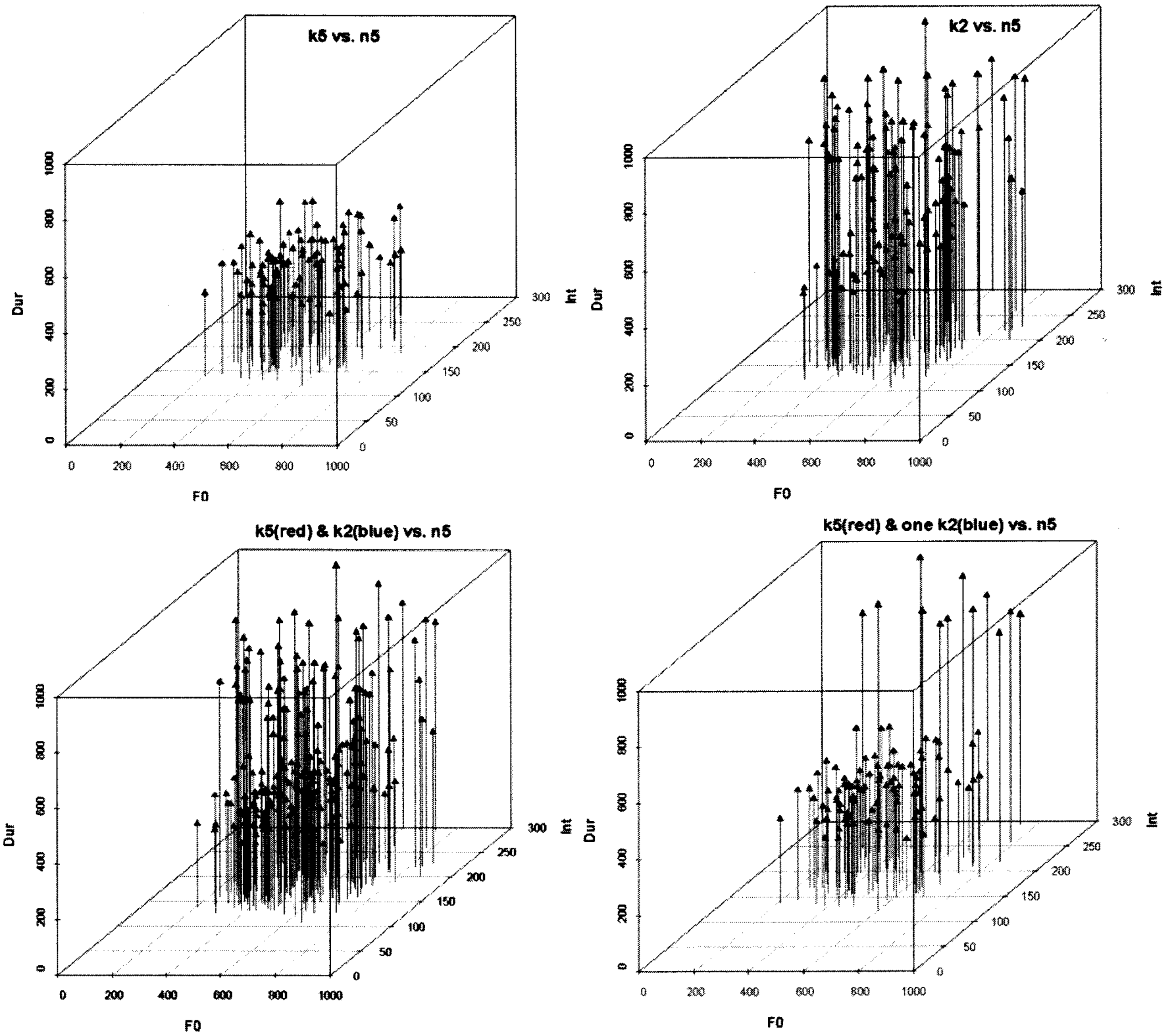


Figure 8. The 3D sentential prosody evaluation models based on evaluation by a Korean phonetician. The points from the learner groups K5' (red) and K2' (blue) are compared with each other. The upper-left panel is for K5' (red) points. The upper-right panel is for K2' (blue) points. The lower-left panel is for K5' (red) and K2' (blue) points. The lower-right panel is for K5' (red) points and a single K2' point (blue).

were 5 for the model native utterances and 4 for all the Korean learners (See <Table 2>). In other words, if the degree of separation among the points in the models are good enough, these models could be used as predictive models for automatic prosody evaluation.

The degree of separation between the N5' and K5' groups (upper-left panel) appears to be weak, meaning that there was little difference in the prosody evaluation between the model native utterances and the learner utterances in the K5' group. Another thing that can be observed is that the N5' points seems more scattered than K5' points. In other words, it can be said that the Korean phonetician was stricter, not necessarily better, in that his "score 5 region" was more compact. The degree of separation between the N5' and K2' (upper-right panel) appears somewhat weak too.

However, if we just compare the learner utterances in groups K5' and K2' in <Figure 8>, we get a clearer picture (upper panels). Noting that all the scales are the same, the degree of separation for K5' points and K2' points is much better than that of any other plots that we have seen.

This pattern can also be observed in the lower-left panel where the two plots are combined. Therefore, these plots can be said to support our initial hypothesis that the consistency of manual prosody scoring is reflected in a prosody evaluation model built from raw physical values of the utterances concerned. As the next step, these models could be used to predict the prosody evaluation scores for other utterance versions of the same sentence. One learner utterance's score set is displayed in the lower-right panel of <Figure 8>. The blue arrowheads represent the semi-automatically evaluated prosody score sets for the Korean learner. Although this score set is originally from K2' group, even if we assume that this set is not from K2' score sets, it can be said to occupy an area near the region occupied by the other K2' score points. Thus, we can give this learner a prosody score of 3 (Refer to <Table 2>).

3.3. Statistical analyses of the prosody evaluation model by a Korean phonetician

The classification table and its confusion matrix from the discriminant analysis are given in <Table 4> and <Table 5> respespively. The two learner utterances from the test group yielded two sets of automatic prosody scores, each of which contained 12 scores. For the test learner from K2' (the right panel of <Table 4>), all the twelve automatic prosody score points were corresply classified as belong<Tabto the K2'. Howe al, among the twelve automatic prosody score points from the test learner of K5', one score point was misclassified as belong<Tabto the other group. Si4>), the other eleven score points were corresply classified, the prhe teive power of the discriminant function seems viable.

When a multiple regression analysis was performed on the three factors, i.e., F0, intensity and duration, they accounted for 42.5% of the manual score points and two factors, i.e. the intensity contour and the segmental durations, were found to be significant predictors for the manual score points (See <Table 6>).

The results from the principal component analysis for all the training dataset are shown in <Table 7> and <Table 8>. <Table 7> shows that the dataset composed of three columns, each of which corresponds to the three prosodic factors, were accounted for by three principal components(PCs). PC1 was able to account for about 47% of the variance of the dataset while PC2 and PC3 were able to account for 32% and 21% respectively. The correlations between the three prosodic factors and the PCs are shown in <Table 8>. PC1 was mainly related with the intensity contour and the segmental durations while PC2 was related negatively with the F0 contour and the segmental durations.

Table 4. The classification table from the discriminant analysis. The number in each cell represents the probability of the automatic prosody score being classified into the predicted group. The left panel is for the test learner from K5' and the right panel is for the test learner from K2'.

predicted **observed**	**K2'**	**K5'**	**total probability**
K5'	0.423926	0.576074	1
K5'	**0.586402**	0.413598	1
K5'	0.144321	0.855679	1
K5'	0.356046	0.643954	1
K5'	0.198097	0.801903	1
K5'	0.247591	0.752409	1
K5'	0.226376	0.773624	1
K5'	0.181814	0.818186	1
K5'	0.343508	0.656492	1
K5'	0.286556	0.713444	1
K5'	0.161533	0.838467	1
K5'	0.203144	0.796856	1
K2'	0.973646	0.026354	1
K2'	0.997847	0.002153	1
K2'	0.995902	0.004098	1
K2'	0.994552	0.005448	1
K2'	0.977365	0.022635	1
K2'	0.985658	0.014342	1
K2'	0.986786	0.013214	1
K2'	0.980868	0.019132	1
K2'	0.986836	0.013164	1
K2'	0.988487	0.011513	1
K2'	0.985466	0.014534	1
K2'	0.981090	0.018910	1

Table 5. The confusion matrix for the classification table.

predicted **observed**	**K2'**	**K5'**	**total**
K2'	12	0	12
K5'	1	11	12
total	13	11	24

Table 6. Results of linear regression analysis on the three factors.

Adjusted R square	0.4247	
Variables (differences of)	Beta	Significance(95%)
(1) F0 contour	-0.0009827	No
(2) intensity contour	0.0071663	Yes
(3) segmental durations	-0.0046342	Yes

Table 7. Eigenvalues from the principal component analysis of the training dataset. PC stands for principal component.

Value	PC1	PC2	PC3
Eigenvalue	1.41	0.95	0.64
% of Variance	46.95	31.64	21.41
Cumulative %	46.95	78.59	100

Table 8. Component loadings from the principal component analysis of the training dataset. PC stands for principal component.

Variable	PC1	PC2	PC3
F0 contour	0.420	-0.900	0.118
intensity contour	0.804	0.141	-0.578
segmental durations	0.766	0.346	0.542

4. Conclusion

The purpose of this paper was to verify the hypothesis that the consistency of the manual prosody evaluation is reflected in a 3D prosody evaluation model whose axes correspond to the three physical properties of the prosody; the F0 contour (x-axis), the intensity contour (y-axis), and the segmental durations (z-axis).

Utterances from the native speakers of English and Korean learners were used to build the prosody evaluation model. All the utterances were evaluated by native teachers of English and a Korean phonetician. Utterances that were given particular scores were selected including those from the native speakers of English, which acted as the model native utterances. All the utterances were semi-automatically evaluated in terms of the three prosodic properties. The automatic evaluation was performed by compr sng each of the utterances with all the select model native utterances. Thus, each utterance produced as many prosody scores, in the form of three-valued coordinates, as the number of model native utterances. The coordinates were then plotted in a 3D prosody evaluation space. The second set of models from a Korean phonetician appeared to support the hypothesis.

It is not possible to directly compare the two sets of models because the sentences were different and there was no control for the segmental proficiency evaluation in the former set of models. However, assuming that there was a uniform or same degree of influence from the segmental proficiency evaluation, it could be said that the scoring consistency of the evaluators was weaker for the native teachers of English than for a Korean phonetician. Considering the fact that the native teachers of English were not trained phoneticians, the criteria for the evaluation of the overall proficiency scores may not have been as balanced as those of the phonetician in between the segmental and prosodic aspects of an utterance. Further experiments could only verify this conjecture.

The second set of models from a Korean phonetician seems viable enough to be used as models against which other utterance versions of the Set B sentence could be prosodically evaluated. If the viability of the models were verified with considerable amount of learner utterances, more such models for other sentences could be built. The significance of this paper is that it is indeed possible to obtain automatic prosody evaluation models by using existing utterances from both the native speakers of English and Korean learners.

Various methods could be used for the automatic prosody evaluation of utterances. Any type of clustering or automatic classification can be used to group known observations and assign an unknown observation to existing groups. By using such standardized distance metrics as the z-scores or the Mahalanobis distance [2], one could evaluate the prosodic aspects of an utterance by classifying a new utterance of the same sentence into one of the existing score groups. For example, one can calculate the Mahalanobis distance of the test utterance to each of the score groups in the model and classify the test utterance as belonging to a particular score group for which the Mahalanobis distance is minimal. The new utterance can be given the particular score for its prosody evaluation. As was done in this work, one can also perform a discriminant analysis and build discriminant functions for the prosody evaluation models and use them to predict the prosody score group of an unknown utterance version of the target sentence.

Although current study examined the relationships between the three prosodic factors in the training dataset by performing a regression and a principal component analyses, one thing that needs to be further investigated is the extent to which each of the three physical properties of the prosody influences the overall prosodic proficiency evaluation. The second set of models from this work mostly showed differences in the z-axis of the segmental durations. There could be ways to improve the technique proposed in [7] so that the differences in the F0 or intensity contour comparison are amplified by using, for example, weights. Certain parts of an utterance, e.g. emphasized, stressed or focused words, could be given more weights in the automatic comparison process. Proper psychological experiments to determine the weights might increase the degree of separation among the 3D score points in the models.

References

[1] Boersma, P. (2001). "Praat, a system for doing phonetics by computer", *Glot International,* Vol. 5(9/10), pp. 341-345.

[2] Mahalanobis, P. C. (1936). "On the generalized distance in statistics", *Proceedings of the National Institute of Science of India* 12, pp. 49-55.

[3] Moulines, E. & Charpentier, F. (1990). "Pitch synchronous

waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication* Vol. 9, pp. 453-467.

[4] Ramus, F., Nespor, M & J. Mehler. (1999). "Correlates of linguistic rhythm in the speech signal", *Cognition* 73, pp. 265-292.

[5] Rhee, S., Lee, S., Lee, Y. & Kang, S. (2003). "Design and construction of Korean-Spoken English Corpus (K-SEC)", *Malsori* 46, pp.159-174.

[6] Yoon, K. (2007). "Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody", *Journal of the Modern British & American Language & Literature* 25(4), pp.197-215.

[7] Yoon, K. (2009). "Synthesis and evaluation of prosodically exaggerated utterances", Phonetics and Speech Sciences, 1(3), pp.73-85.

• 윤규철 **(Yoon, Kyuchul)**
영남대학교 영어영문학부
경상북도 경산시 대동 214-1
Tel: 053-810-2145 Fax: 053-810-4607
Email: kyoon@ynu.ac.kr
관심분야: 음성학, 음운론
현재 영어영문학부 교수