An Acoustical Study on the Syllable Structures of Korean Numeric Sounds

Byunggon Yang*

ABSTRACT

The purpose of this study was to examine the syllable structures of ten Korean numeric sounds produced by ten students. Each sound was normalized by its maximum intensity value and divided into onset, vowel, and coda sections after finding abrupt or visible changes in energy values or cumulative values of lower spectral energy at each pulse point using four Praat scripts. Then, segmental durations and cumulative intensity values of each syllable were obtained to find a statistical summary of the syllable structure. Intensity values at 100 proportional time points were also collected to compare the ten sounds.

Results showed as follows: Firstly, there was not much deviation from the grand average duration and intensity for the majority of the sounds except the two diphthongal sounds on which their boundary points varied among the speakers. Secondly, the onset point for the CV or CVC category sounds and the boundary between the vowel and the nasal or lateral sound were easy to identify, which may be automatically traced later. Thirdly, there seems some tradeoff among the sections maintaining the same total duration per each syllable. Further studies on syllables with various onsets or codas would be desirable to make a general statement on the Korean syllable structure.

Keywords: Korean numeric sounds, syllable, spectral processing, temporal organization

1. Introduction

A syllable plays an important role in the speech synthesis and analysis. However, the syllable is said to be easy to identify but impossible to define(Sloat, Taylor & Hoard, 1978; Gigerich, 1992; Roach, 1999). Many people can tell how many syllables a given utterance have immediately after listening to them. There might be just a small difference in the number of syllables when the speakers have different dialects. The syllable can be divided into three parts: the onset, the nucleus or peak, and the coda (McMahon, 2002). The most prominent part is called the peak, which usually consists of a vowel or resonant consonant with higher sonority. Both the onset and the coda consist of consonants and they are optional constituents. The segmentation of its constituents is another challenge in doing further research on the temporal

* Department of English Education, Pusan National University

microstructure of speech at the level of phoneme (Port, Al-Ani & Maeda, 1980; Port, Dalby & O'Dell, 1987).

In order to find some reliable criteria for speaker identification, Yang(2001) analyzed the acoustic parameters of nine Korean numbers and synthesized and modified them to make a perceptual test on the sameness of a pair of model signals followed by a modified sound. He found that the subjects perceived the same sound quality within the range of 6.6 dB of intensity variation, 10.5 Hz of pitch variation, and 5.9% of the first three formant variation. Yang & Kang(2002) proposed a speaker identification method using difference sum and correlation coefficient calculated from a pair of intensity level matrices of band-pass-filtered numeric sounds. They obtained matrices of five intensity levels at 100 proportional time points. Even though those studies could not provide a reliable method for speaker identification, his use of cumulative energy from the narrow band spectra seems to be applicable to the segmentation of vowel or consonant sections within a syllable. In other words, the spectral energy measured at onset, peak, or coda sections in a syllable may vary. This study will pursue the segmentation with that idea. Yang(2006) compared English CV syllables produced by five American speakers in a controlled context. He segmented each syllable by a few visible acoustic criteria. For the syllable with onsets such as stops or affricates, he selected the first pulse of the syllable from the Praat edit window and decided the nearest zero-crossing point as a boundary between its onset and peak. For /la/, he searched the highest intensity peak in the syllable and moved backward to reach a cycle without the usual three smaller peaks followed by each pulse point of the vowels /a/. His segmentation method seems to include some human errors in exactly pinpointing the syllable boundary. In other words, visual selection of a time point from a small edit window of Praat may invoke some inconsistency.

In this paper the author attempts to determine boundaries of the onset, peak and coda sections of the ten numeric sounds using the following acoustic information of each syllable: energy values and sums of spectral energy along the pulse points of a given sound. Then, temporal structures and spectral curves of the ten Korean numeric sounds produced by ten Korean students will be examined. Those numeric sounds display various syllabic structures in Korean and are widely examined in the field of speech engineering. Results of this study may be useful in speech synthesis and analysis. With the detailed analysis of the syllable structure of numeric sounds, speech synthesis may use segments faithful to each segment rather than diphones which are derived from simply dividing a two-syllable word in the middle of each syllable. Moreover, this study may contribute to the comparison of the syllable structure of a native language and that of the learners' production of a foreign language, which will lead to remedy their chronic pronunciation problems.

2. Method

2.1 Speech data and subjects

The speech data were chosen from the phonetically balanced speech database collected by SiTEC at Wonkwang University. The data were originally recorded using Senheizer HMD224X on a digital audio tape at a sound-proof booth. Then, the speech sounds were digitized at a sampling rate of 16 kHz and the resolution of 16 bits on a KAY CSL 4300B. The author used 110 numeric sounds produced by ten students. All of the students were 21 years old and born and spent their lives in Seoul.

2.2 Stimuli

Korean numbers are produced either descended from the original Korean sound or based on the Chinese pronunciation(Yang, 2001). The speech data listed only the sounds based on the Chinese. The ten numeric sounds in Korean can be classified into four categories according to the syllable structure. Sounds '2'(/i/) and '5'(/o/) consist only of a vowel(V). Sounds '4'(/sa/) and '9'(/ku/) are composed of an onset followed by a vowel(CV). Sound '1'(/il/) has a vowel followed by a coda(VC). The remaining sounds '0'(/jəŋ/), '3'(/sam/), '6'(/jʊk/, '7'/tʃil/), '8'(/pal/) and '10'(/sip/) have a vowel preceded and followed by a consonant(CVC). After examining the speech data, the author found that only 3 out of the 10 subjects produced the sound '6' with the final plosive burst at the end. It was difficult to pinpoint the transition from the glide /j/ to the vowel /ʊ/. The same problem occurred in the sound '0'. However, the transition from the vowel to the coda showed an abrupt noticeable gap in those two sounds. The script for the sound '1' was used to analyze those sounds and tentatively grouped and discussed. For the sound '10', no cases were found with the burst. The script for the sound '4' was used to measure its structure and grouped with the sounds '4' and '9' in this paper. The numeric sounds using the same script were grouped for discussion.

2.3 Segmentation and analysis

Four Praat scripts to process the sounds of the four categories were developed to facilitate the segmentation of each component section of a syllable. Each script generally included three procedures. Firstly, the onset and offset points of each sound were traced from the energy values of the sound and normalized to its maximum intensity value(Yang, 2006) and the preceding and following segments were set to zero to avoid any interference in the measurement of marginal intensity values. Secondly, the boundary was decided between the onset and the vowel peak or that between the vowel peak and the coda mostly referring to the printed information of energy values and cumulative spectral energy sums at each pulse point.

The spectral energy was summed up from 141 Hz to 1704 Hz every 30 Hz interval. The CV boundary was set to half duration of the first two pulses backward from the first pulse point of the sound. The script drew a possible boundary point in the edit window finding the mean point of the energy sums between the vowel and the onset or the coda. When the author accepted the point or clicked a modified point, the script automatically selected the nearest pulse point as the boundary. That way, more consistent boundary decisions along the pulse points were made possible. Thirdly, a statistical summary of acoustical measurements and their percentage proportion of each syllable component was obtained. The summary included the duration or intensity of each syllabic segment, ratios of the segment to the total duration or cumulative intensity values. <Figure 1> illustrates the segmented portions of the numeric sound '3' produced by a subject. The shaded area indicates vowel section preceded by the fricative /s/ and followed by the nasal /m/. One can notice that the sound was normalized and those segments preceding and following the syllable were set to zero.

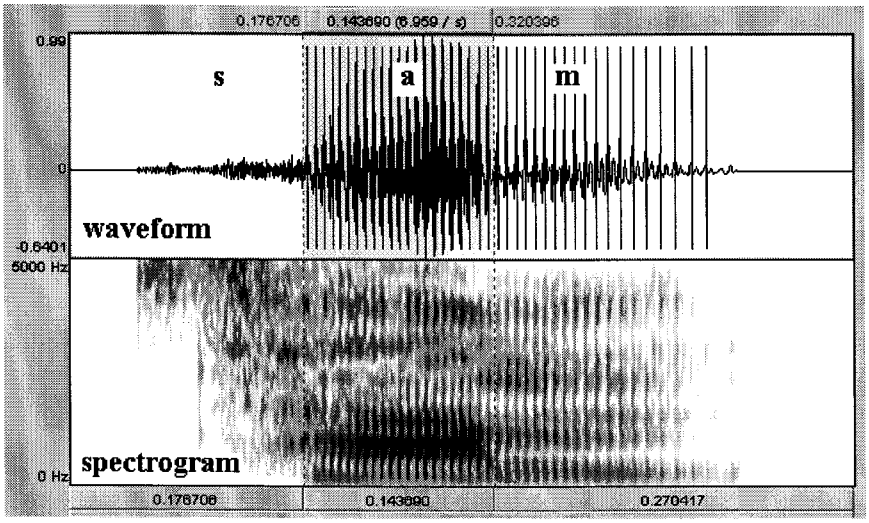

Figure 1. An example segmentation of the sound '3'

3. Results and Discussion

3.1 Sounds '2' and '5'

<Table 1> shows a statistical summary of them. The sound '2' was produced slightly longer than the sound '5'. <Figure 2> illustrates the average intensity curves of them. The grand average duration of the ten sounds was 332 ms. The grand average intensity of the ten sounds was 80 dB. The duration and intensity of the two sounds fall on the grand average. One can

note that the two curves show quite a similar pattern. There is only a negligible shift from the 37th time point with the same intensity level. The curve starts around 60 dB and reaches the energy peak at one third of the total duration, and then its intensity slowly goes down.

Table 1. The average duration and intensity values of the sounds '2' and '5'. ave stands for average; sd, standard deviation; dur, duration in milliseconds; db, intensity level in dB.

Values	'2' ave	'2' sd	'5' ave	'5' sd
Voweldur	342	53	329	51
Voweldb	7764	205	7777	217
dbave	81	2	81	1
dbsd	8	1	7	2

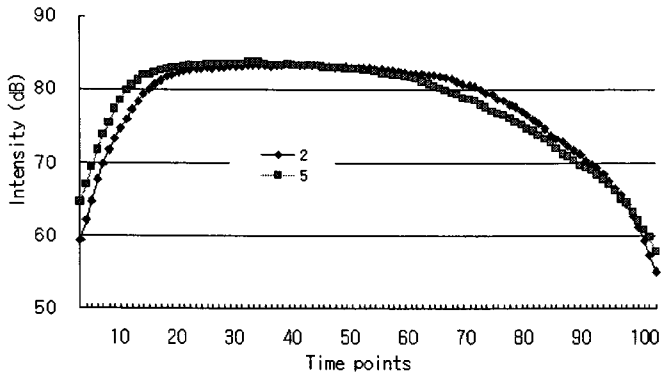

Figure 2. The average intensity curves of the sounds '2' and '5'

3.2 Sounds '4', '9' and '10'

<Table 2> shows a statistical summary of the sounds '4', '9' and '10'. <Figure 3> illustrates the average intensity curves of them. According to the table, the total duration of the sound '4' amounts to 387 ms, which is a little longer than that of the sound '9'. The duration of the sound '10' is the shortest which may be related to the final stop /p/. Its total duration will match the grand average duration if the final silence period before the invisible stop burst is included. The fricative onset durations for the sounds '4' and '10' are 119 ms and 159 ms, respectively. The total intensity sums of the three sounds are similar. The dips for the boundary between the onset and the vowel appear the shortest for the sound '9' followed by the sound '4'. The frication noise for the sound '10' comes out the longest. From the ratio data, one can see that 69% to 78% of the syllable accounts for the vowel sections of the sounds '4' and '10' while the vowel for the syllable '10' is 45%, less than half of the total duration. The curve of the sound '9' starts at 68 dB and descends slightly to 66 dB at the 14th time point and ascends again for the following vowel. Its duration from the onset to the dip (75 ms) can be

regarded as the voice onset time for the stop sound /k/. The dip of the sound '10' comes around the 48th time point. Since the vowel duration ratio is 55%, the vowel section must start at the 55th point. Therefore, the dip may not always be interpreted as the onset of the vowel.

Table 2. The average duration and intensity values of the sounds '4', '9' and '10'. ave stands for average; sd, standard deviation; Ons, onset; dur, duration in milliseconds; db, intensity level in dB; R, ratio in %.

Values	'4' ave	'4' sd	'9' ave	'9' sd	'10' ave	'10' sd
Onsdur	119	23	75	16	159	36
Voweldur	267	57	267	54	128	25
CVdur	387	64	341	55	286	47
Onsdb	1981	449	1493	347	4055	419
Voweldb	5240	454	6128	500	3549	510
CVdb	7221	154	7620	240	7605	144
OnsdurR	31	6	22	5	55	6
VoweldurR	69	6	78	5	45	6
OnsdbR	27	6	20	5	53	6
VoweldbR	73	6	80	5	47	6
dbave	78	1	81	2	80	1
dbsd	9	1	9	1	8	1

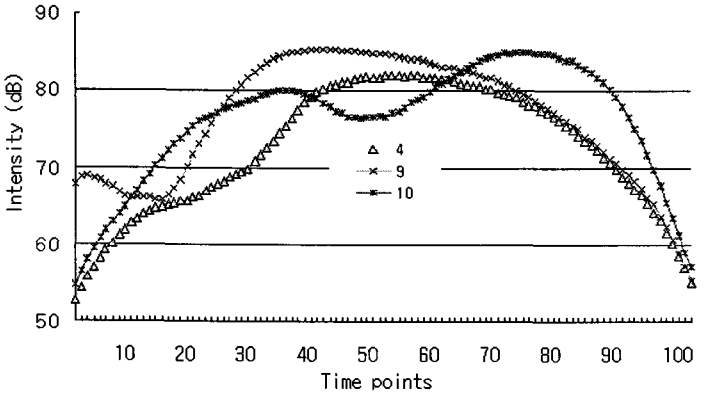

Figure 3. The average intensity curves of the sounds '4', '9' and '10'

3.3 Sounds '0', '1', and '6'

<Table 3> shows a statistical summary of the sounds '0', '1', and '6'. <Figure 4> illustrates the average intensity curves of them. From the table, one can note that the total duration of the sound '6' with the stop sound /k/ comes out the shortest. The other two sounds have only a difference of 17 ms. For the sound '6', there were only four cases with the plosive burst at the

end of the syllable. Their silence gap ranged from 17 ms to 51 ms. As was seen in the discussion of the sound '10', the total duration for the sound '6' may be comparable to the other two sounds if the silence gap is added to its total duration. Therefore, there must be some interaction among the duration of the component sections in the syllable structures. The speakers might have produced the sounds keeping a certain total duration of a syllable constant.

Table 3. The average duration and intensity values of the sounds '0', '1', and '6'. ave stands for average; sd, standard deviation; Ons, onset; dur, duration in milliseconds; db, intensity level in dB; R, ratio in %.

Values	'0' ave	'0' sd	'1' ave	'1' sd	'6' ave	'6' sd
Voweldur	171	43	114	28	146	26
Codadur	168	52	207	39	89	40
VCdur	339	30	321	47	235	41
Voweldb	4021	991	2897	668	5046	856
Codadb	3401	1004	4798	494	2816	922
VCdb	7422	165	7695	293	7862	206
VoweldurR	51	13	36	7	63	10
CodadurR	49	13	64	7	37	10
VoweldurR	54	13	37	8	64	11
CodadbR	46	13	63	8	36	11
dbave	79	1	81	2	82	2
dbsd	8	1	8	2	7	1

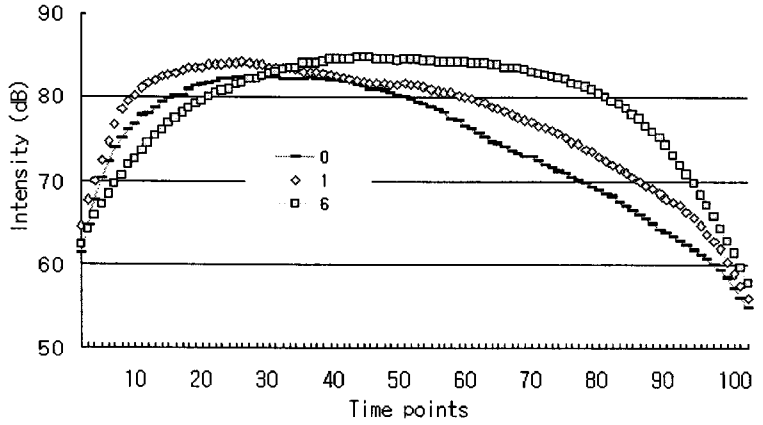

Figure 4. The average intensity curves of the sounds '0', '1', and '6'

The standard deviations are not very big for the duration measurements but there are huge deviations for the sounds '0' and '6', which may be derived from the continuously moving glides

of the diphthongs. Since it was quite difficult to pinpoint the boundary for the two diphthongal sounds, the average point of the spectral energy sums was used here. The intensity sum of the coda for the two sounds shows around 1000 dB. Normally the segmental deviations are around 500 dB for the majority of the ten numeric sounds examined here. The deviations of the total syllable range from 165 to 293 dB. For the sound '1', the vowel section falls on two-thirds of the syllable. The boundary for the coda /l/ looks quite clear. <Figure 5> illustrates the section /l/ of the sound '1' by a subject. The energy drop along the pulse points can be used to choose the boundary.

From <Figure 4>, the energy curve for the sound '0' with the nasal /ŋ/ degraded faster than the other two sounds. Its average intensity comes the lowest because of the nasal sound with antiresonances (Picket, 1987:77, 124). If the peak point of the sound '6' is slightly moved backward along the time point, it will fit to the curve '1' except the abrupt drop of the sound '6' near the offset point.

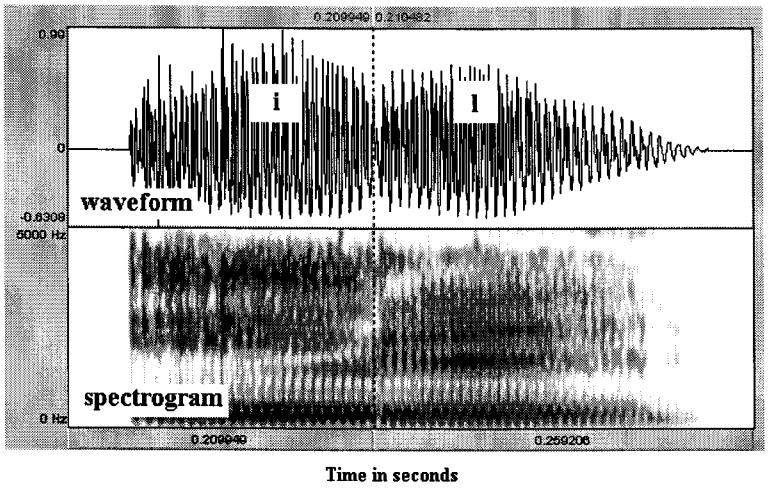

Figure 5. An example segmentation of the sound '1'

3.4 Sounds '3', '7' and '8'

<Table 4> lists a statistical summary of the sounds '3', '7' and '8'. <Figure 6> illustrates their average intensity curves. The total duration of the three sounds ranges from 334 to 391 ms. Their spectral energy sums of the sounds '7' and '8' are similar but that of the sound '3' with the nasal coda is the lowest as was in the case of the sound '0'. The standard deviations are not very big in the duration measurements. It may be related to the clear boundary for the sound '3' as is seen in <Figures 1 and 3>.

From <Figure 6>, the energy peak of the sound '3' is at the 51st time point. Those for the sound '7' and '8' are at the 50th and 47th points, respectively. Since the vowel arrives at the

middle points of those syllables, their shapes show somewhat of the same durations for a nicely divided share of the total syllable. The syllable weight for the VC portions seems heavier than that of the CV ratio due to the sonorant sound /l/. Interestingly, the speakers produced the coda /l/ with very high energy. Here again, the normal values of the standard deviations indicate that the boundary decisions were made rather consistent. Further studies on the comparison of the syllable structure would be quite interesting in a detailed study on the Korean syllable structure after recording the same vowel with various onsets or codas (i.e., sa, sam, san, sal, etc.) which is beyond our discussion here.

Table 4. The average duration and intensity values of the sounds '3', '7', and '8'. ave stands for average; sd, standard deviation; Ons, onset; dur, duration in milliseconds; db, intensity level in dB; R, ratio in %.

Values	'3' ave	'3' sd	'7' ave	'7' sd	'8' ave	'8' sd
Onsdur	120	25	114	17	61	16
Voweldur	119	35	87	34	102	30
Codadur	152	23	151	42	171	43
CVdur	239	29	200	38	163	31
VCdur	271	51	237	58	273	57
CVCdur	391	48	351	54	334	53
Onsdb	2014	586	2519	495	1260	436
Voweldb	2431	553	2007	570	2445	519
Codadb	2690	220	3160	623	3744	675
VCdb	5122	586	5168	637	6189	556
CVdb	4445	195	4526	588	3705	533
CVCdb	7135	186	7687	228	7449	172
OnsdurR	31	7	33	7	19	7
VoweldurR	30	7	24	7	30	7
CodadurR	39	2	43	8	51	8
CVdurR	61	2	58	8	49	8
VCdurR	69	7	67	7	81	7
OnsdbR	28	8	33	7	17	6
VoweldbR	34	8	26	7	33	7
CodadbR	38	3	41	8	50	8
CVdbR	62	3	59	8	50	8
VCdbR	72	8	67	7	83	6
dbave	77	2	80	1	79	2
dbsd	9	1	7	2	8	1

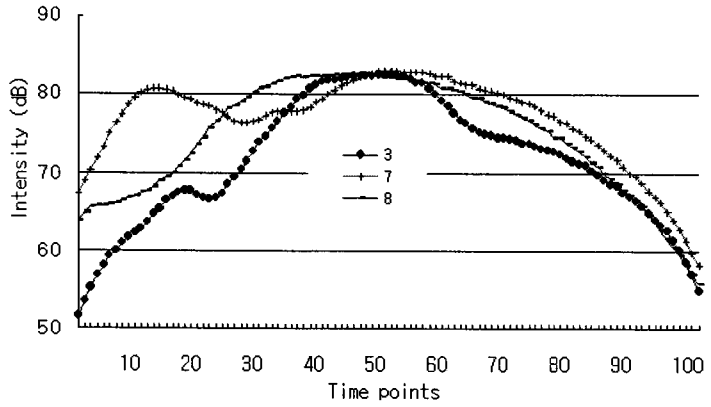

Figure 6. The average intensity curves of the sounds '3', '7', and '8'

4. Conclusion

This study examined the syllable structures of ten Korean numeric sounds from the clearly recorded speech database. Each sound was segmented to find the onset, vowel and coda sections using the Praat scripts. The energy values and the spectral energy sums on the pulse points were determined to provide some information on the possible section boundary. Temporal and spectral summary on each syllable structure was presented and discussed along with the average intensity curves of the ten sounds.

Results showed as follows: Firstly, the grand average duration and intensity of the ten sounds were 332 ms and 80 dB respectively. There was not much deviation from the grand average values for the majority of the sounds except the two diphthongal sounds on which the boundary decision was not very consistent. Secondly, the onset point of the sounds under the CV or CVC categories and the boundary between the vowel and the nasal or lateral sound were easy to identify, which may be automatically traced in the future study. Thirdly, there seems some tradeoff among the sections maintaining the same total duration per each syllable. Those sounds that ended with the stop sound had relatively shorter duration. From those results, one may conclude that the spectral average sum or energy values on the pulse points may be possible cues to segment the syllable. However, the high standard deviations for the diphthongal sounds may require more consistent boundary selection. Further studies on syllables with various onsets or codas would be desirable to make a general statement on the Korean syllable structure.

References

Gigerich, H. J. 1992. *English Phonology: An Introduction.* Cambridge: Cambridge University Press.

Lee, E. 2003. "A comparative study of syllable structures between French and Korean in real utterances." *Speech Sciences* 10(2), 237-248.

Lee, O. & Kim, J. 2005. "Syllable-timing interferes with Korean learners' speech of stress-timed English." *Speech Sciences* 12(4), 95-112.

McMahon, A. 2002. *An Introduction to English Phonology.* Oxford: Oxford University Press.

Roach, P. 1999. *English Phonetics and Phonology.* Cambridge: Cambridge University Press.

Picket, J. M. 1987. *The Sounds of Speech Communication: A Primer of Acoustic phonetics and Speech Perception.* Austin, Texas: Pro-ed.

Port, R. F., Al-Ani, S. & Maeda, S. 1980. "Temporal compensation and universal phonetics." *Phonetica* 37, 232-252.

Port, R. F., Dalby, J. & O'Dell, M. 1987. "Evidence for mora-timing in Japanese." *Journal of the Acoustical Society of America* 81, 1574-1585.

Sloat, C., Taylor, S. H. & Hoard, J. E. 1978. *An Introduction to English Phonology.* Englewood Cliffs, N.J.: Prentice-Hall International, Inc.

Yang, B. 2001. "Speaker variation in number production by males." *Speech Sciences* 8(3), 93-104.

Yang, B. & Kang, S. 2002. "A study on speaker identification by difference sum and correlation coefficient of intensity levels from band-pass filtered sounds." *Speech Sciences* 9(3), 3-16.

Yang, B. 2006. "An acoustical study of English CV syllables." *Speech Sciences* 13(4), 127-140.

Yun, I. 2004. "Temporal variation due to tense vs. lax consonants in Korean." *Speech Sciences* 11(3), 23-36.

received: January 30, 2007
accepted: March 13, 2007

▲ Byunggon Yang
English Education Department, Pusan National University
30 Changjundong, Keumjunggu, Pusan, 609-735, Korea
Homepage://fonetiks.info/bgyang
Tel: 010-9618-7636
E-mail: bgyang@pusan.ac.kr