Perceptual Experiment on Number Production for Speaker Identification[1)]

Byunggon Yang
Dongeui University

ABSTRACT

Acoustic parameters of the nine Korean numbers were analyzed by Praat, a speech analysis software, and synthesized by SenSynPPC, a Klatt formant synthesizer. The overall intensity, pitch and formant values of the numbers were modified dynamically by a step of 1 dB, 1 Hz and 2.5% respectively. The study explored the sensitivity of listeners to changes in the three acoustic parameters. Twelve male and female subjects listened to 390 pairs of synthesized numbers and judged whether the given pair sounded the same or different. Results showed that 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. The male and female groups showed almost the same perceptual ranges. Also, an asymmetrical structure of high and low boundary was observed. The ranges may be applicable to the development of a speaker identification system while the method of synthesis modification may apply to its evaluation data.

Key words: perception, synthesis, Korean numbers, speaker identification

1. Introduction

People produce numbers frequently in everyday lives. Each person has a series of unique numbers for identification. Therefore, it is natural and easy for an individual to produce numbers to identify himself or herself in a web-based business transaction. Because most personal computers are equipped with sound-input and -output systems, it will be desirable to make most of the system to record each individual's voice and analyze it to match future voice inputs for speaker identification. According to Hirahar and Kato(1992), the absolute formant frequencies will provide cues to speaker identification whereas the relative differences among the formants can be employed to identify vowels. However, since the formant values partially represent the speech output in the source-filter model (Fant, 1960), the amplitude and pitch information should be included to correctly identify the speaker among tens of thousands of the customers registered. Especially, the number production by the same speaker will not always be the same acoustically so that it may be difficult to find what identifies the individual among several possible candidates. Even though human perception in different people may not be more accurate than the machine comparison, it may be meaningful to investigate how human beings process the sound difference at the first stage. In other words, the machine comparison of acoustic parameters can be almost indefinite in their combination if we are using acoustical data with all the digits below zero. One way to solve the problem may be by collecting a large speech database and extracting some unique statistical patterns of individual differences. Another way may be pursued by a perceptual experiment to find a certain range of the same sound quality. The author believes that a successful machine identification will be just a little more sensitive within the range of human discrimination. The perceptual results may provide some insight into where we should focus during the identification procedures by computers. Besides, the synthesis method may be applicable for training the machine and evaluating whether it will correctly identify the synthesized pair with a gradual increment or decrement of the parameters. Sometimes we cannot obtain enough data to train computers for all the possible sets of human speech.

Previous studies on vowel perception (Yang, 1995:142, Table 5) revealed that there were certain formant ranges of the same vowel quality. The first formant varied for almost 200 Hz unnoticed by the listener. Also, the F2 range came out around 400 Hz and that of F3 around 800 Hz. Basically, the listener showed the wider range of the same sound quality for the higher formant values. The range became wider with diphthongs (Yang, 1996). The uniformly modified diphthongs led to comparable perceptual ranges with those of the monophthongal study. The result reflected the psychoacoustical characteristics of critical bands (Zwicker, 1962). In those experiments,

*This work was supported by Korea Research Foundation Grant (KRF-99-041-A00010).

only one formant value was modified at a time. Even though subtle variation within the vocalic segment was automatically implemented to synthesize more natural outputs, the formant values were not varied dynamically across time. In this study, the amplitude, pitch, and formant values will be modified dynamically.

Korean numbers are produced in two different ways: *hana, duul, set...* (the original Korean sound) or *il, i, sam...*(based on the Chinese pronunciation). This study will deal with the latter set which are short and more widely used. '0' is realized by either /jʌŋ / or /koŋ/. Here /jʌŋ/ for '0' will be used to capture the diphthongal variation. The ten numbers consist of single syllables. If we look at their syllable structure, the peak vowel /i/ is used in the numbers '1', '2' and '7'. Also, the vowel /a/ is at the center of '3', '4' and '8'. Numbers '5' and '9' are produced with rounded lips. Two diphthongs in '6' and '0' show dynamic variation across each syllable. If we examine the consonantal structure, a fricative, a stop and an affricate are used at the onset of '4', '8' and '7'. Two nasals /m, ŋ / and a lateral /l/ occur as codas including a stop /p/ in '6'. The author tried to synthesize the Korean numbers in a pilot study and found that acoustic parameters for the onset and coda varied so greatly that this perceptual study will limit the observation of intensity and pitch perception to the resonant segments modified. However, the formant values will be modified throughout the syllable.

The aim of this study is to explore the perceptual range of the same sound quality of the synthesized Korean numbers. For this purpose, we will analyze pitch, amplitude and formant variation from the author's production of the nine Korean numbers except '8'. Next, after synthesizing the numbers, each acoustic parameter will be modified dynamically. Then, each pair of synthesized numbers will be randomly presented to listeners to find out the sensitivity of listeners to changes in the amplitude, pitch and formant values. Also, the gender difference in the perception will be examined.

2. Method

2.1 Subjects

The author decided to produce the nine numbers and to check the naturalness of the synthesized ones because the extraction of synthesis parameters required fine recording and an accurate evaluation of the synthesized sound. Six male and six female students participated in a listening session at Dongeui University. The subjects were students of Dongeui University and all had normal hearing and health. They marked "the same" for more than eight out of the ten pairs of the same sounds. Their ages ranged from 21 to 25. The average age for the females was 21, while that of the males was 24. The average height of the male subjects was 175 cm while that of the females was 164 cm.

2.2 Procedures

2.2.1 Analysis of numeric sounds

The author read the numbers clearly at a slow rate. The input samples were digitized at a sampling rate of 22 kHz on a G3 PowerPC notebook. The average duration of the numeric syllables was 714 ms with an SD of 150 ms. /juk/ has the shortest duration of 356 ms excluding the coda. /sa/ has the longest duration of 872 ms including the fricative onset. Each number was dynamically analyzed by *Praat*, a speech analysis software. Amplitude, pitch, formant frequency values and zero-crossing rates were gathered from computer estimates and stored on a disk with the synthesis header file. Acoustic parameters were drawn on a spectrogram to visually double-check their validity. Pitch was collected every 5 ms applying weights to choose the best candidate. The global pitch average and standard deviation were employed to avoid any leap of pitch values across the syllables, which might be impossible in the human articulatory movements. Smoothed amplitude variation was determined within the range of 100 Hz. The average value within a 10-ms window replaced the pitch values out of the one standard deviation from the average. To determine the noise characteristics of the onset and coda, zero-crossing rates within the same window were determined for AH and AF values. If the pitch value was greater than 0, the automatic formant extractor algorithm(sl) was used while the value was 0 or undefined, the algorithm(burg) was employed. The bandwidth values of the voiceless section were assigned to the parameter AF.

2.2.2 Synthesis

SenSynPPC, a Klatt formant synthesizer was employed to import the synthesis file. Any obvious errors were immediately corrected to make better synthesized sounds. Each sound was stored on the computer at a sampling rate of 20 kHz. The major acoustical parameters such as amplitude, pitch and the first four formant values and zero crossing rates varied dynamically throughout the syllable segment. The number '8', /pal/ was excluded because it was difficult to synthesize naturally enough at this stage. Its onset had mixed acoustic features of burst, frication and aspiration, which could not be determined by the zero-crossing rate. The others sounded quite natural enough to be considered as the author's voice by four males and females. The author derived the appropriate bandwidths (B1 to B3) from Fant's equations (Fant 1972:47). The bandwidth affects the intensity and quality of the neighboring formants. The others were set to the default values of the Klatt synthesis file. Once the reference file was established, each acoustical parameter was varied dynamically using a spreadsheet software. Intensity and pitch were increased or decreased by a step of 1 dB or 1 Hz, respectively. Figure 1 shows the pitch and amplitude variation. The upper thicker line indicates the pitch model while the lower one shows the amplitude model.

Figure 1. The amplitude and pitch variation for '0', /jʌŋ/. The two thick lines indicate the amplitude or pitch of the model. The numbers after the plus or minus sign are the increment or decrement unit in dB of the parameters.

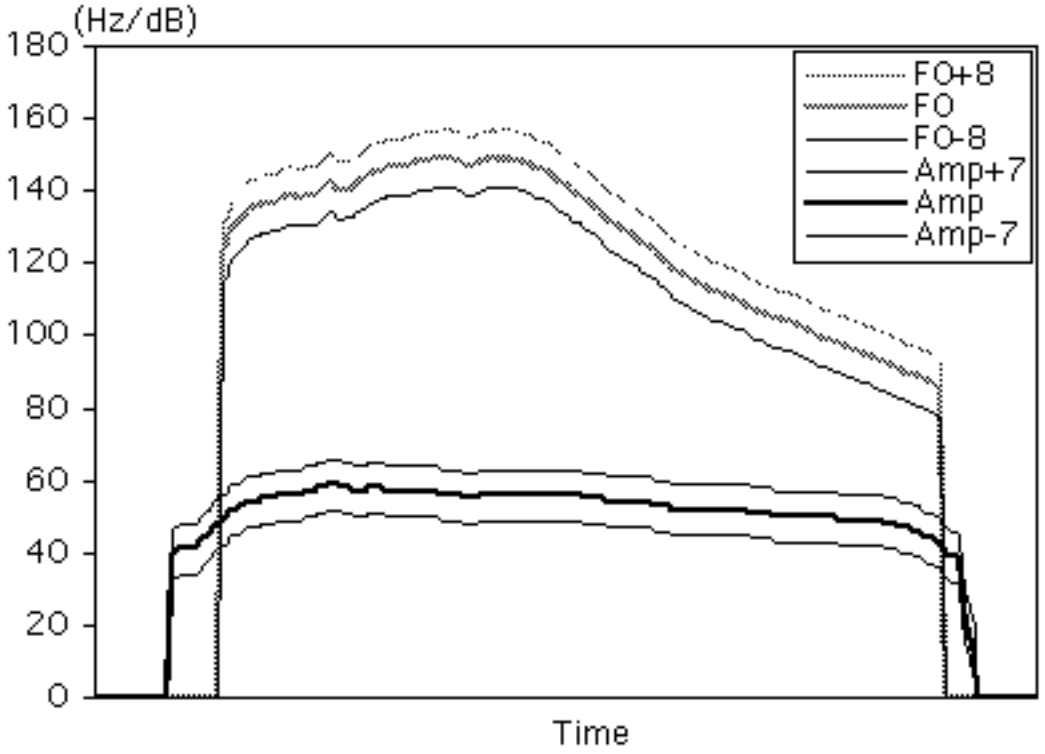
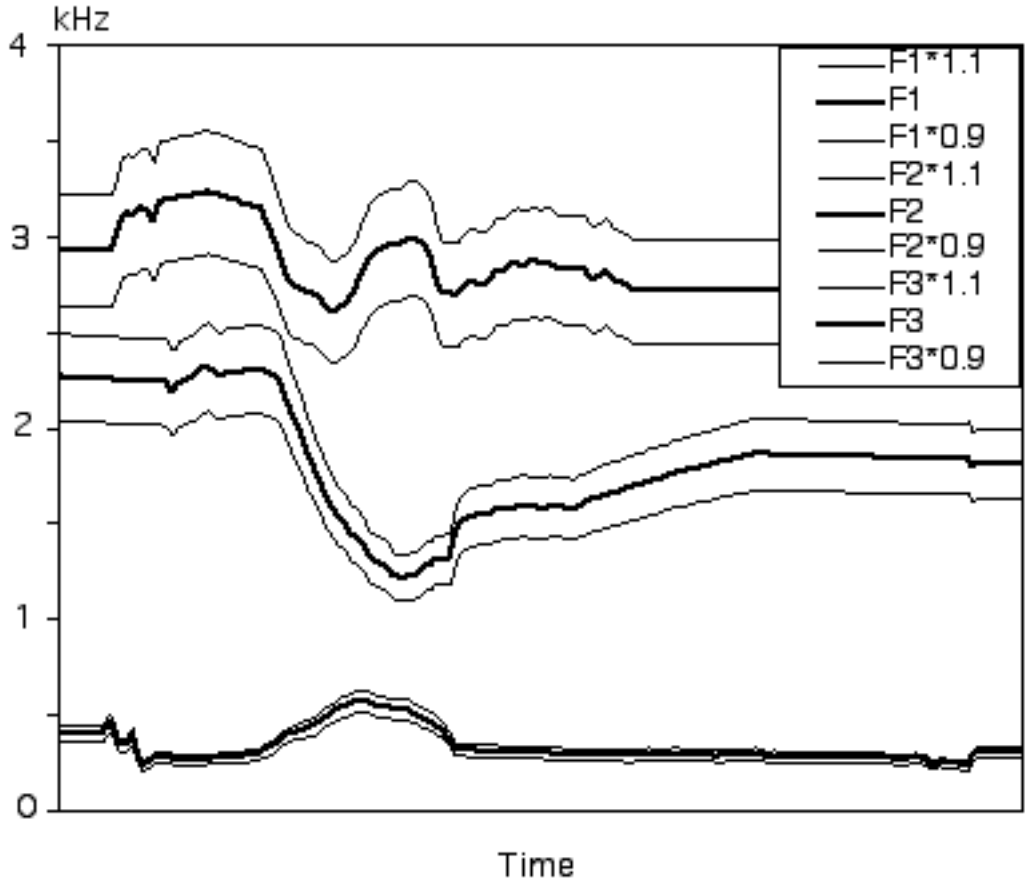


Figure 2 shows the formant variation in eight steps. Here the first three formant values were simultaneously modified so that the whole syllable quality changed. The changes were done proportionally by a step of 2.5 percent up or down from the model. The upper or lower boundary was set to 11% because the vocal tract ratios of female to male were 0.89 for Swedish (Fant, 1975), 0.89 for Dutch (van Nierop, Pols and Plomp, 1973; Pols, Tromp and Plomp, 1973), 0.86 for English (Peterson and Barney, 1952) and 0.82 for Korean (Yang, 1996). More than 11% may lead to a gender shift of the sound quality. This way eight pairs of stimuli for each number were included in the perception test.

2.3 Listening

The listening samples consisted of the nine Korean numbers in 390 randomized pairs. The subjects judged whether each pair sounded the same or different on a given item of an answer sheet.

Figure 2. The formant variation for '0', /jʌŋ/. The three thick lines indicate the first three formants of the model. The numbers after the asterisk are the increment or decrement ratio in percent of the parameters.



Five practice pairs followed by the model pairs were given so that they could adjust to a comfortable listening level and the 10 pairs with the same sounds were randomly presented in the test to check the validity of each subject's response. The 12 subjects discriminated more than 80% of them correctly. Every pair of the stimuli was numbered in English and the author's voice was played to help them locate the correct item on the answer sheet. The subjects circled the given number during the response time of 1200 ms if they perceived the same sound pair, otherwise they used a slash mark. The listening session was conducted in a language laboratory with headphone sets. The session lasted for 30 minutes.

The answers were marked on a page in the order of variation to find a certain range of the same sound quality for each number. When a subject circled a stimulus pair across a slashed stimulus on the same parameter, the value of the middle pair was included in the same sound range. However, a circled pair after two slashed stimuli from the model was discarded. This kind of interpolation may reflect the subtle perceptual range of each subject.

3. Results and Discussion

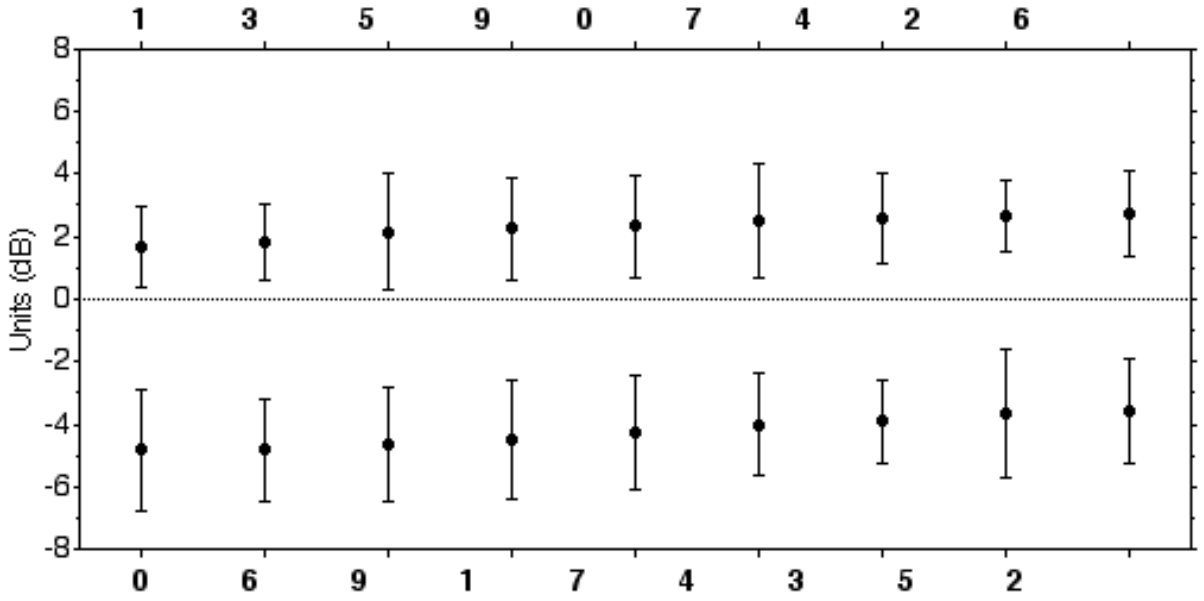
3.1 Intensity Perception

Table I shows the average value of high and low boundary of intensity perception by the male and female groups. The final row indicates the overall average of the 12 subjects. According to the Table, the gender difference in each number seems not very large. The average difference of the low boundary is 0.87 dB while that of the high boundary is 0.43 dB. Therefore, the two groups were merged together to observe any numeric trends. Figure 3 shows the average and one standard deviation of the high and low boundary of intensity perception of male and female groups together. The numeric variation seems quite minimal. The lowest value of the high boundary is -5.3 dB for the number '0' while that of the high boundary is 3.3 dB for the number '6'. The values of the high boundary tend to be lower than those of the low boundary. The average standard deviation among the low boundary is 1.77 dB while that of the high boundary is 1.49 dB, which indicates quite stable results across the numbers.

Table I. The average boundary of intensity perception by the male (M) and female (F) groups. "lo" after each number denotes the low boundary while "hi" does the high boundary. "Ave" describes the overall average of the 12 subjects for each number.

	0lo	**0hi**	**1lo**	**1hi**	**2lo**	**2hi**	**3lo**	**3hi**	**4lo**	**4hi**	**5lo**	**5hi**	**6lo**	**6hi**	**7lo**	**7hi**	**9lo**	**9hi**
M	-4.3	2.2	-4.2	1.7	-3.2	2.5	-4.3	2.2	-3.7	2.5	-3.0	2.0	-4.7	2.2	-3.7	2.3	-4.2	2.0
F	-5.3	2.5	-4.8	1.7	-4.0	2.8	-3.5	1.5	-4.3	2.7	-4.3	2.3	-5.0	3.3	-4.8	2.7	-5.2	2.5
Ave	-4.8	2.3	-4.5	1.7	-3.6	2.7	-3.9	1.8	-4.0	2.6	-3.7	2.2	-4.8	2.8	-4.3	2.5	-4.7	2.3

Figure 3. The average and one standard deviation of the high and low boundary of intensity perception of the 12 subjects. The numbers are indicated above (high boundary) or below (low boundary) the frame, in bold face.



3.2 Pitch Perception

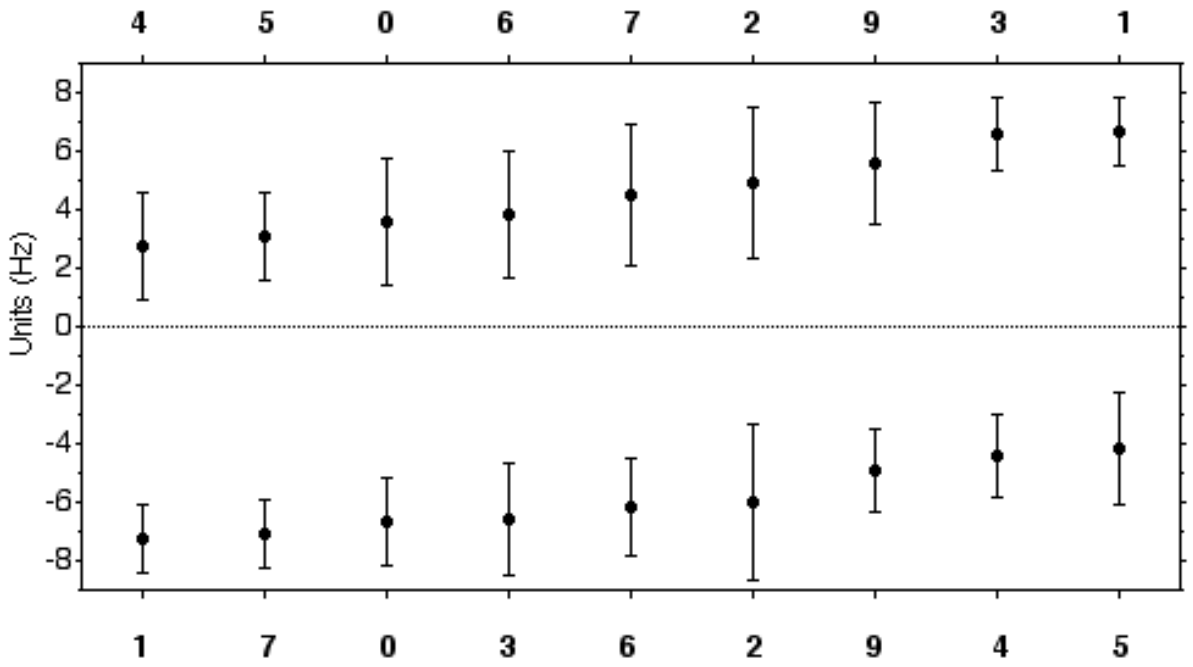
Table II lists the average boundary of pitch perception by the male and female groups. The final row indicates the overall average of the 12 subjects. The average gender difference of the low boundary is 0.61 Hz while that of the high boundary is 1.3 Hz. Those values are almost as negligible as in the intensity perception. The two groups are merged together to observe any numeric trends. Figure 4 illustrates the average and one standard deviation of pitch perception. Some numeric variation can be observed. '0', '2' and '9' are placed at the

same place in the sorted order of the average boundary. However, '1', '4' and '5' are in the opposite place of the high and low boundary. The lowest value of the low boundary is -7 Hz for the number '3' while the highest of the high boundary is 7 Hz for the number '1'. The average standard deviation among the low boundary is 1.96 Hz while that of the high boundary is 2.32 Hz.

Table II. The average boundary of pitch perception by the male (M) and female (F) groups. "lo" after each number denotes the low boundary while "hi" gives the high boundary. "Ave" describes the overall average of the 12 subjects for each number.

	0lo	**0hi**	**1lo**	**1hi**	**2lo**	**2hi**	**3lo**	**3hi**	**4lo**	**4hi**	**5lo**	**5hi**	**6lo**	**6hi**	**7lo**	**7hi**	**9lo**	**9hi**
M	-6.5	2.3	-6.8	7.0	-6.3	5.8	-7.0	6.3	-4.7	3.3	-4.0	3.7	-5.7	3.5	-7.0	4.2	-4.5	4.3
F	-6.8	4.8	-7.7	6.3	-5.7	4.0	-6.2	6.8	-4.2	2.2	-4.3	2.5	-6.7	4.2	-7.2	4.8	-5.3	6.8
Ave	-6.7	3.6	-7.3	6.7	-6.0	4.9	-6.6	6.6	-4.4	2.8	-4.2	3.1	-6.2	3.8	-7.1	4.5	-4.9	5.6

Figure 4. The average and one standard deviation of high and low boundary of pitch perception of the 12 subjects. The numbers are indicated above (high boundary) or below (low boundary) the frame, in bold face.



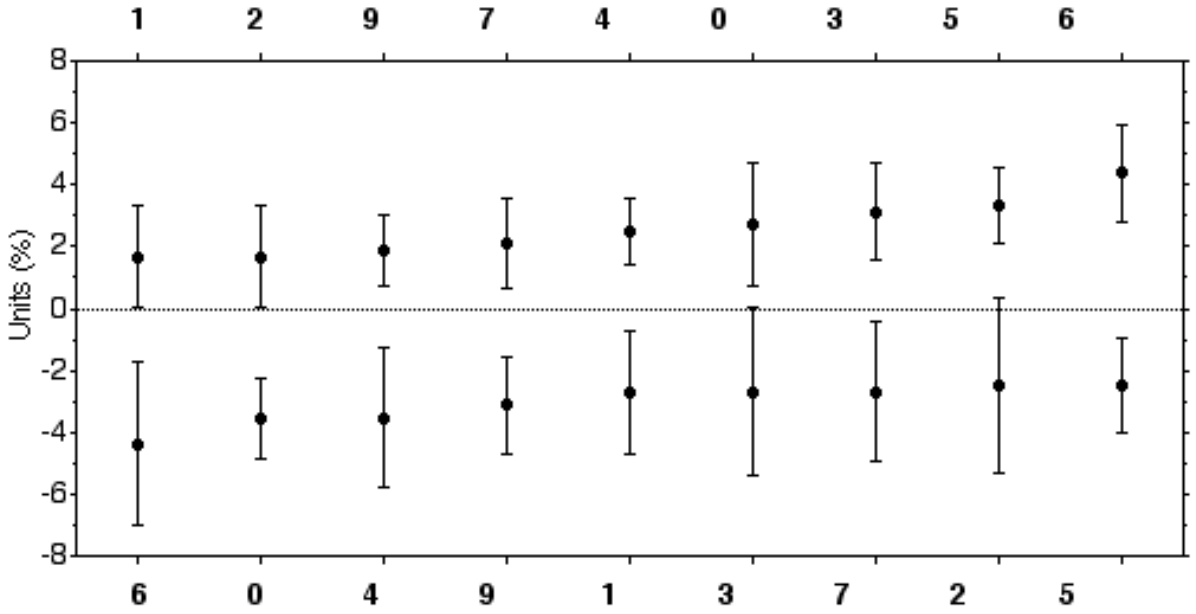
3.3 Formant Perception

Table III describes the average boundary of formant perception by the male and female groups. The final row indicates the overall average of the 12 subjects. Here again, the gender difference seems not very large. The average difference of the low boundary is 1.1% while that of the high boundary is 1%. Figure 5 depicts the average and one standard deviation of formant perception. Here again, the numeric variation seems quite small. The numbers '3, 4, 6' have 3.8% while the number '6' reaches 4.6%. The average standard deviation among the low boundary is 2.17% while that of the high boundary is 1.7%.

Table III. The average boundary of formant perception by the male (M) and female (F) groups. "lo" after each number denotes the low boundary while "hi" reflects the high boundary. "Ave" describes the overall average of the 12 subjects for each number.

	0lo	0hi	1lo	1hi	2lo	2hi	3lo	3hi	4lo	4hi	5lo	5hi	6lo	6hi	7lo	7hi	9lo	9hi
M	-2.9	2.1	-2.1	2.5	-1.7	2.5	-1.7	3.8	-3.3	2.9	-2.5	2.9	-3.8	4.6	-2.1	1.7	-3.3	1.7
F	-4.2	3.3	-3.3	0.8	-3.3	0.8	-3.8	2.5	-3.8	2.1	-2.5	3.8	-5.0	4.2	-3.3	2.5	-2.9	2.1
Ave	-3.5	2.7	-2.7	1.7	-2.5	1.7	-2.7	3.1	-3.5	2.5	-2.5	3.3	-4.4	4.4	-2.7	2.1	-3.1	1.9

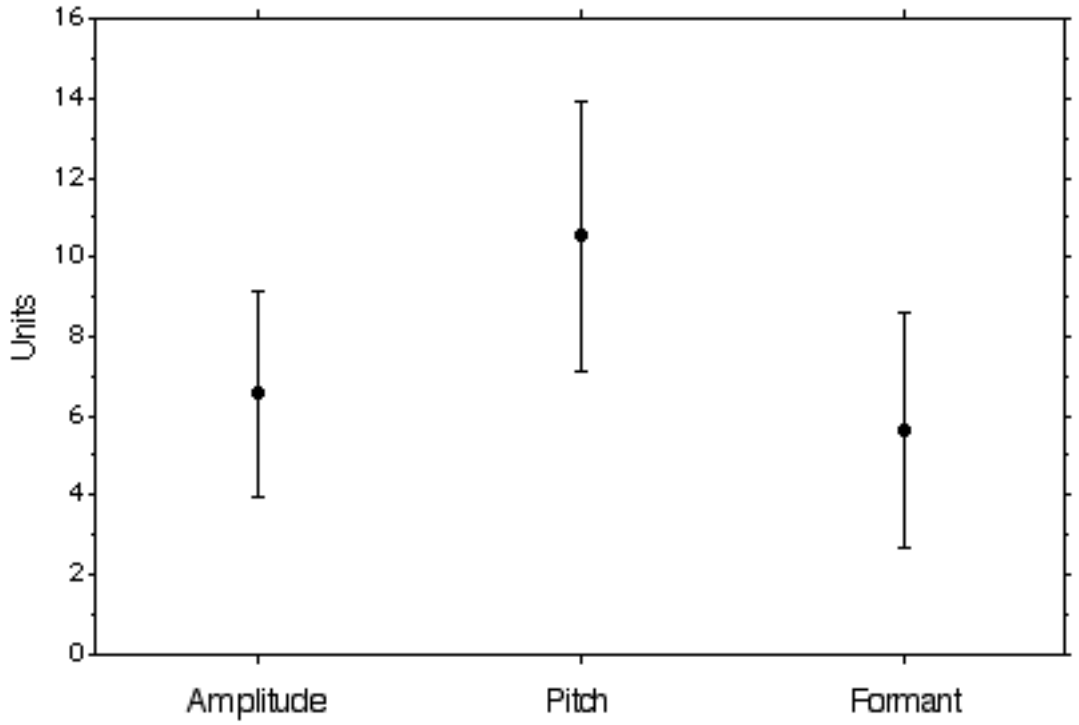
Figure 5. The average and one standard deviation of high and low boundary of formant perception of the 12 subjects. The numbers are indicated above (high boundary) or below(low boundary) the frame in bold face.



3.4 Discussion

The ranges between the lower to higher boundary were determined by adding the high and low boundary values after removing the minus sign. Figure 6 illustrates the average range and one SD of amplitude, pitch and formant perception together. The global average range of amplitude for the 12 subjects was 6.6 dB with an SD of 2.6 dB. That of pitch was 10.5 Hz with an SD of 3.4 Hz. The formant variation ratio amounted to 5.9% with an SD of 3%. Typical values of the just noticeable difference (JND) in amplitude for a pure tone or wide-band noise were in the range of 0.3 to 1.0 dB (Riesz, 1928; Miller, 1947).

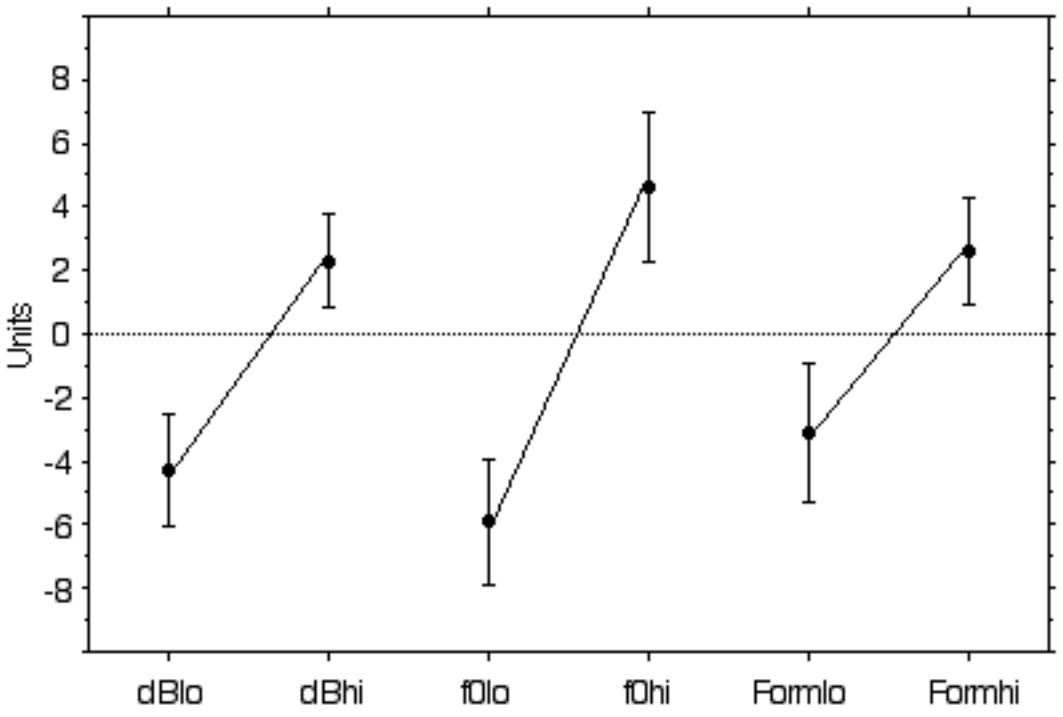
Figure 6. The global average range and one SD of amplitude, pitch and formant perception.



Since we experimented on the speech stimuli, the amplitude range might become wider. However, the average amplitude range seemed a rather narrow one. This might be related to the syllable structure of the Korean numbers. Most numbers had wider resonant segments, thus the overall change in amp litude and pitch influenced the discrimination task. The pitch range also seemed to reflect the sensitivity of human hearing. Stevens (1998) reported the JND for normal listening levels tended to be constant at about 1 Hz for frequencies up to about 1 kHz. The range became larger when the frequency of the stimulus tone increased. He also pointed out that the amplitude of the tone influenced the JND level, which might lead to different ranges for each number. Stevens (1952) found out that the JND in the frequency of the single spectral prominence, like a vowel with a single formant, came out in the range of 10 to 20 Hz. This value might be comparable to the 6% range of formant perception. In other words, 6% of F1 with 300 Hz will be around 18 Hz. This value is substantially lower than the ranges in the formant discrimination of Korean monophthongs (Yang, 1995). It might be related to the dynamic changes of the first three formant values simultaneously. The sound quality must have been changed greatly.

To investigate whether the high and low boundary showed any different perceptual pattern, the author calculated the boundaries separately. Figure 6 shows the average range with one standard deviation bar. The overall range might be roughly the distance between the low average and high average. Interestingly, most subjects allowed a wider range in the negative side in which the model sound was followed by the sound with lower amplitude or pitch. However, formant variation range, up or down the model sound, came out quite comparable. It might be related to an asymmetrical structure of human perception. Stevens (1998:238) reported a considerable asymmetry in the tuning curves for the high frequency probe tones. As the frequency of the probe tone went higher, a sharp rise occurred. On the other hand, a gradual rise occurred below the frequency of the tone.

Figure 7. The average range and one SD of high and low boundary of amplitude ("dB"), pitch ("f0") and formant perception ("Form"). "lo" indicates the low boundary while "hi" does the high boundary.



Another way of interpreting this result was to consider masking effects even though the two stimuli were not presented in an overlapping manner. Listeners might be influenced by the interaction of the two stimuli when they responded to the following stimulus with higher amplitude or pitch in a short time interval of 100 ms. For non-simultaneous masking, the amount of masking depended on the time interval between the masker and the test stimulus (Stevens, 1998). Future studies on the asymmetrical nature may be possible by presenting the stimuli pair in the opposite order.

The lowest range for intensity came out 1 for intensity and 0 for pitch. The highest range amounted to 14 dB for intensity and 16 Hz for pitch. Only one subject showed the possible 14 dB range in the intensity perception while two subjects marked 2 dB difference in their intensity ranges. There were eight cases of the 16 Hz range for the pitch perception, which suggests the necessity of further expansion of the range in the perceptual test even though this extreme range was marked by a few subjects. Five male or female subjects scored at least one pair of zero formant range. This zero range actually means at least 5% allowance because the modification was done every 2.5% step up or down the model.

Finally, the correlation coefficients among the three parameters were determined using the StatView+, a statistics software. Table IV lists the coefficients. Individual comparison suggests somewhat independent discrimination patterns.

Table IV. Correlation coefficient matrix for amplitude, pitch and formant variation.

	Amplitude	Pitch	Formant
Amplitude	1		
Pitch	0.231	1	
Formant	0.294	0.166	1

4. Conclusion

The author analyzed the acoustic parameters of the nine Korean numbers by Praat, and synthesized them by SenSynPPC and modified them dynamically. The 12 male and female subjects listened to 390 pairs of the synthesized numbers in a quiet laboratory room. The model number was presented first and followed by a modified one and rated whether the given pair sounded the same or different. Results showed 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 variations. The male and female groups showed almost the same range. The low standard deviation indicated stable response patterns across the numbers. In the perception of pitch and amplitude, an asymmetrical pattern of high and low boundary was observed. Most subjects allowed a wider range in the lower boundary which might be related to the masking or perceptual asymmetry of frequency selectivity on the probing tone.

Some of the subjects responded the same sound quality even after the pitch changed by 8 Hz, which was the upper or lower limit in this study. The average range may go higher if one adopts the pitch variation above 8 Hz difference. The fluctuation of subjects' attention during the 30 minute session might have an affect on the result. Separate studies on pitch and intensity perception would lead to more accurate results.

Future studies will be made on the stress perception of synthesized sounds with dynamic variation of pitch. Also, these synthesized files with subtle variation may be applicable for training the computer to refine the speaker identification system.

Acknowledgment

This work was partially supported by a grant from the Interdisciplinary Research Program (Contract No. 1999-2-302-106-5) of the KOSEF.

References

Hirahara, T. and H. Kato. 1992. The effect of F0 on vowel identification. In Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistics Structure*. pp. 89-112. Tokyo: Ohmsha Publishing.

Fant, G. 1960. *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton.

Fant, G. 1975. Speech Production, *STL-QPSR*, 2-3, pp. 1-19.

Miller, G.A. 1947. Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *Journal of the Acoustical Society of America*, 19, pp. 609-619.

Peterson, G. E. and Barney, H. L. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24, pp. 175-184.

Pols, L. C. W., Tromp, H. R. C. and Plomp, R. 1973. Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America*, 53, pp. 1093-1101.

Riesz, R. R. 1928. Differential intensity sensitivity of the ear for pure tones. *Physical Review* 31, pp. 867-875.

Stevens, K. 1952. *The Perception of Sounds Shaped by Resonant Circuits*. Sc. D. dissertation. Cambridge, MA: The MIT Press.

Stevens, K. 1998. *Acoustic Phonetics*. Cambridge, MA: The MIT Press.

van Nierop, D. J. P. J., Pols L. C. W. and Plomp, R. 1973. Frequency analysis of Dutch vowels from 25 female speakers, *Acustica*, 29, pp. 110-119.
Yang, B. 1995. A perceptual study of synthesized Korean monophthongs. *Korean Journal of Linguistics*, 20-3, pp. 127-146.
Yang, B. 1996. A perceptual study of Korean diphthongs synthesized. *Korean Journal of Linguistics*, 21-3, pp. 829-843.
Yang, B. 1996. A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics*, 24, pp. 245-261.
Zwicker, E. 1962. Subdivision of the audible frequency range into critical bands. *Journal of the Acoustical Society of America*, 33. p.248.

Received: January 1, 2001.
Accepted: February 28, 2001

Byunggon Yang
English Department, Dongeui University
24 Kayadong, Pusanjingu, Pusan, 614-714, Korea
Tel:+82-51-890-1227, FAX: +82-51-890-1209
e-mail: bgyang@hyomin.dongeui.ac.kr