Vowel Perception by Formant Variation

Byunggon Yang

Department of English Language and Literature, Dongeui University, 24 Kayadong, Pusan 614-714 South Korea

Abstract: The acoustic parameters of six Korean vowels produced by a healthy male were analyzed to elaborate
synthesize the six vowels by a formant synthesis method. Then, F1 and F2 values of the synthesis file were modified
by 50 Hz while F3, by 100 Hz over and under the original values but not interfering adjacent formants to obtain 270
stimuli. Twenty male and females listened to the original vowel followed by the corresponding synthesized one and
judged whether they sounded qualitatively 'similar' or 'different'. Results showed that males responded 'similar' within
the range of variance by an average of 163 Hz for F1; 415 Hz for F2; 843 Hz for F3. Females indicated 40 Hz higher
range than those of males. The ranges supported the nonlinearity of human auditory scale to the speech sound.

INTRODUCTION

Human beings can perceive various aspects of sound which include loudness, pitch, length, and timber. Recently
many studies were made to clarify complex auditory scales of the human ear. Sone and phon are examples of
loudness scales, while mel and bark, pitch scales. Though those scales were derived from subjective experiments
with a small number of subjects, they clearly showed a coherent trend of nonlinearity in human tonal perception
once we look at those scales against a linear frequency axis. Nonlinear Phon scale indicates tone of lower frequencies
should be louder to be perceived than that of higher frequencies. Also, the ear seems most sensitive around 3,400 Hz
on the scale, which equals to a quarter wave resonance of the human ear 2.5 cm long. Some scale was proposed to
catch nonlinear loudness which can be linearly comparable. Fant(1) proposed a pitch scale called technical mel which
can be determined by a logarithmic nonlinear scale. Critical band or bark (2) also indicates nonlinear jump in the
higher frequencies and higher resolution in the lower frequencies. Moreover, the cochlea has a spatial frequency
mapping which clearly depicts non-linearity of the ear(3). The basal portion of the basilar membrane is sensitive to
higher frequencies while the apical portion responds to lower frequencies. Almost half of the cochlea is assigned to 3
kHz which may mean higher resolution in speech like signals. The other half reacts to the frequencies above.

However, questions may arise whether the same scale may apply to the complex human voice. Most studies on
the auditory scaling were done by using simple tonal sounds. Therefore, this study will focus on vowel perception
by formant variation using a sophisticated speech synthesizer. Holmes(4) reported that he could make an almost
perfect replica of human speech by using a formant synthesis method. This study is important because once we
capture the human auditory scale or the boundary of speech perception for each vowel, then we may apply the range
to automatic speech recognition. It will also contribute greatly to the understanding of speech perception.

METHOD

The acoustic parameters of six Korean monophthongs produced by a healthy male subject with normal hearing
were analyzed to synthesize the six vowels /a, e, i, o, u, / and by a formant synthesis method until each
corresponding synthesized vowel was perceived as almost the same as the naturally produced vowel. These
peripheral vowels were selected because they do not show much dialectal variation in Korean. Speech inputs were
made by SoundEdit at 22 kHz sampling rate. F0 was collected by an autocorrelation method at every 5 ms.
amplitude envelopes were used to collect relative amplitude values. Four formant values of each vowel were
collected from the signal using Signalyze 2.45. Each file was synthesized using SenSyn1.0 on Macintosh LC630
until they sounded like those of the original sounds by repeatedly modifying synthesis parameters of the file. The
output sampling rate was set to 20,000 Hz per second. Then, F1 and F2 values of the synthesis file were modified
by a step of 50 Hz over and under the original values but not interfering adjacent formants fixed. Thus, F1 of the
vowel [i] varied within the range of 200-950 Hz. F2 varied 500 Hz over and under the original values. F3 varied 500
Hz over and under the original values by a step of 100 Hz. This way 270 synthesized vowel stimuli were obtained.
A total of 20 male and female students attending Dongeui University participated in perceptual tests in a quiet room.
Random stimuli were played by a Macintosh computer on a pair of speakers at a comfortable level. Each subject
listened to the original vowel followed by the synthesized one in 0.5 sec and judged whether they sounded
qualitatively 'similar' or 'different'. They marked 'similar' or 'different' on an answer sheet numbered. All the marks
were collected to determine the upper and lower limit of frequency range within which the subject perceived the
stimuli similar.
RESULTS AND DISCUSSION

Table 1 shows the frequency range of each vowel within which the subjects perceived each pair of stimuli similar. The letter m followed by each formant number means male while f means female. The first number under each formant indicates the lower limit of the range while the second one denotes its higher limit. An average range of higher limit minus lower limit comes under the table.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>F1m</th>
<th>F2m</th>
<th>F3m</th>
<th>F1f</th>
<th>F2f</th>
<th>F3f</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>720-910</td>
<td>1080-1540</td>
<td>2300-3200</td>
<td>720-950</td>
<td>1040-1490</td>
<td>2310-3190</td>
</tr>
<tr>
<td>i</td>
<td>260-420</td>
<td>2330-2830</td>
<td>3030-3400</td>
<td>220-410</td>
<td>2290-2970</td>
<td>3000-3500</td>
</tr>
<tr>
<td>o</td>
<td>510-630</td>
<td>700-990</td>
<td>2110-3060</td>
<td>500-730</td>
<td>640-930</td>
<td>2100-3060</td>
</tr>
<tr>
<td>u</td>
<td>300-410</td>
<td>570-1100</td>
<td>2100-3080</td>
<td>300-390</td>
<td>560-990</td>
<td>2100-3000</td>
</tr>
<tr>
<td>A</td>
<td>540-680</td>
<td>1070-1450</td>
<td>2170-3310</td>
<td>500-660</td>
<td>1110-1470</td>
<td>2300-3400</td>
</tr>
</tbody>
</table>

Average range 163 415 843 208 453 882

Specifically, males responded 'similar' to the stimuli within the range of variance by an average of 163 Hz for F1; 415 Hz for F2; 843 Hz for F3. Females indicated 40 Hz higher range than those of males. This result may be related to higher formant values of female vowel production because of shorter vocal tract (5). Roughly, F2 has double the value of F1 while that of F3 amounts to five times that of F1. It implies that any perceptual experiment by a step of less than 50 Hz may not give any significant perceptual difference. F3 range of vowel [i] is 370 Hz because F2 value is already near 2,500 Hz. Any wider range may lead to perceptual confusion. Vowel [e] shows the widest variation in F1 because the distinction between Korean [e] and [e] is being lost. The variation among vowels seems to be related to "sufficient perceptual contrast" by Lindblom(6). He suggested that vowels maintain certain perceptual distance to secure sufficient perceptual contrast. Yang (7) supported the notion in the comparative study of Korean and English vowels normalized. Thus, all the formant values are not overlapped in the perceptual experiment and the center frequency of the range correlates strongly with the original formant value of each vowel (R squared equals 0.996 for male group;0.995 for female group). Male and female groups show similar perceptual range, which supports the notion that subjective perceptual experiment has reliability. Correlation coefficient between male and female ranges amounts to 0.959 in total. Comparing each formant yields 0.867 for F1, 0.544 for F2, and 0.913 for F3. Perceptual range shows fine resolution in the lower frequency range while coarse resolution in the higher frequency range. Validity of exact formant frequency measurement should be tuned to a perceptually appropriate range. In other words, any sophisticated equipment for exactly tracing the formant values would not be useful, especially in the higher frequency region if the human ear may not discriminate such physical difference. Further studies will be desirable to vary three formant values together or in a sentence context. These ranges may be applicable to speech recognition system to transform acoustic data to the auditory scale as the ear does.

REFERENCES