THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.21, No.1E 2002. 3. pp. 31~41

What You Hear is What You See?

Seung-Jae Moon*

*Department of English, Ajou University
(Received November 12 2001; accepted March 10 2002)

Abstract

This study aims at investigating the relationship between voice and the image information carried within the voice. Whenever we hear somebody talking, we form a mental image of the speaker. Is it accurate? Is there a relationship between the voice and the image triggered by the voice? To answer these questions, speech samples from 8 males and 8 females were recorded. Two photos were taken for each speaker: the whole body photo (W) with physical characteristics present, and the face close-ups (F) without much physical details revealed. 361 subjects were asked to match the voices with the corresponding photos. The results showed that 5 males and 5 females (with W) and 2 males and 4 females (with F) were correctly identified. More interestingly, however, even in the mismatches, there was a strong tendency for participants to agree on which voice should correspond to which photo. The participants also agreed much more readily on their favorite voice than on their favorite photo. It seems voice does carry certain information about the physical characteristics of the speaker in a consistent manner. These findings have some bearings on understanding the mechanism of speech production and perception as well as on improving speech technology.

Keywords: *Voice, Image, Speaker indentifcation*

I. Introduction

Whenever we hear somebody talking, over the shoulder or over the phone, we immediately form, in our brain, a certain mental image about the person. This everyday phenomenon raises many questions: How accurate is the image triggered by the voice alone? In other words, does this image truly reflect the owner of the voice? Do different people conjure up a similar image from the same voice?

This kind of questions, which try to relate the acoustic output with some psychological and physiological factors, have not been addressed seriously in academic arena.

Corresponding author: Seung-Jae Moon (moon@madang.ajou.ac.kr)
Ajou University, Suwon, 442-749, Korea

Of course, there are works investigating the mechanical workings of vocal folds and their acoustical characteristics [1]. But little effort was put into the investigation of the characteristics of voice in terms of their psychological values. There is some evidence, though, that researchers start to look into this linguistic and psychological realm[2].

There are, however, two previous pilot studies[3,4] which promised some positive results: they show that there seems to be a definitely positive relationship between the voice and the image triggered by the voice.

These results have several implications. Speech technology has developed considerably. There are many commercially available speech synthesizers. Now imagine what it would be like to have a synthesizer capable of producing different timbers of voices for different situations. The voice of

a synthesizer does not have to be always the same. For example, reading a book to a child, the synthesizer can produce a voice which might inspire a 'mom-type' person. And when it is used for tele-marketing, it will be more efficient with a voice which sounds to be coming from a pleasant person. It would all be possible if we knew what characteristics in the voice make us picture the owner of the voice as we do.

The results also have some strong implications toward theories of speech perception as well, since they will shed some light on the issue of how we make use of non-segmental information in the voice. Everyone knows that voice does much more than just carrying linguistic information. It carries emotion, for one thing. And it is almost impossible to hear a friend's voice without picturing his/her face. These phenomena need to be investigated to understand human perception of voice.

But the previous studies mentioned have some limitations to draw a definite conclusion. For example, the experimental setting used in Moon (1999)[4] was less than ideal. For the perception experiment in the study, many subjects (thirty people at a time) were asked to watch a picture projected on the wall by a beam projector from a scanned picture while listening to the speaker playing the recordings of each speaker three times in a row. In this condition, the resolution of the projected pictures was much less clear than the actual photograph. And the subjects did not have any control over how many times they listened to the tape. And the number of subjects participated (100) was somewhat limited.

The present study attempts to further the systematic understanding of the visual and physical information carried in the voice by expanding Moon (1999)[4].

II. Perception Experiment

To address the issues mentioned above, a perception experiment was carried out in which subjects matched different voices with different photos.

A. Speakers

There were 8 male and 8 female Korean speakers. They were chosen from a group of middle- and high-school teachers in Korea who were at the time attending a special seminar at the author's university. The ages of male speakers were 32, 32, 33, 34, 35, 35, 38, and 39. And those of female speakers were 26, 28, 30, 31, 34, 35, 35, and 40. (For the privacy of speakers, these ages are given in an ascending order, not in the order of speaker numbers; therefore, speaker *m1*'s age is not necessarily 32.) Care was taken to choose an age-homogenous group as closely as possible to eliminate the possibility of voices revealing age differences. However, there were still some age variations (especially for female speakers) because there were not many volunteers. (Many people were very reluctant after learning that not only their voices but also their photos were going to be recorded. The author was able to get 16 volunteers out of 128, after a continuous and long series of persuasion.)

B. Speech Materials[1)]

Speech of the total duration of about 16 to 23 seconds was recorded for this experiment. To eliminate any possible revelations of the speaker characteristics other than voice, speakers were asked to read the same section of a Korean fairy tale. Each speaker was seated in front of an Electro-Voice 635 dynamic omni-directional microphone at about 7" distance. The recording was made with TASCAM PA-1 DAT machine (at 48 kHz sampling rate). No special instruction was given except to speak normally.

Each recording was later digitized into a separate file using Kay Elemetrics CSL 4300B at 20 kHz sampling rate with an 8 kHz low-pass filter.

C. Photographs[1)]

Two photographs were taken for each subject using Minolta Maxxum 7000 camera with 35-105 mm zoom lens: one a whole-body photo (W henceforth), and the other face-only close-up (F henceforth). This was to determine

1) Speech materials and the pictures are the same as used in the previous study[4].

whether physical characteristics influence the judgement as discussed above. Whole-body photos were taken from a fixed distance with a fixed zoom and the speakers stood right in front of a large board which served as a kind of reference frame. Face close-ups were taken in such a way that only the face was shown without any telltale revelation

Figure 1. Male Whole-body photos (Black and white version).

Figure 2. Female Whole-body photos (Black and white version).

of his/her body.

Each photograph was printed on a regular 3.5"x5" photograph paper. Then these photographs were digitally scanned at 720 dpi resolution and edited. To ensure that their relative physical characteristics were intact for whole-body photos, each person's photo was adjusted so that the reference board would be exactly the same size for all speakers. For face-only close ups, the photos were resized so that their heads were approximately the same size. Then these 8 scanned images were randomly aligned on one letter-size paper (in a portrait orientation) and printed on a high-resolution color laser printer. These photos are shown in figures 1 through 4. Although the photos used in the experiment were color, these figures are black and white and reduced at on appropriate rate to make presentation in the easier.

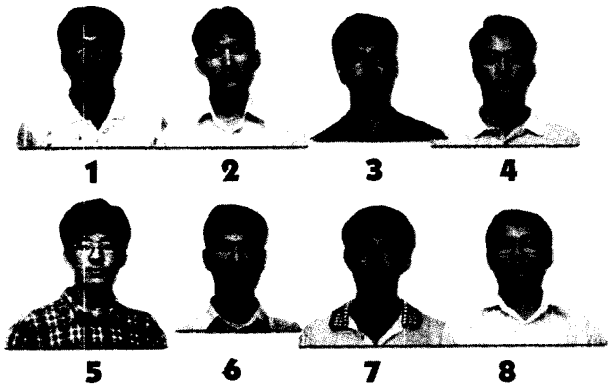

Figure 3. Male Face-only photos (Black and white version)

Figure 4. Female Face-only photos (Black and white version)

D. Subjects

Subjects were recruited among undergraduate students at Ajou University. (None of the subjects participated in the previous study[4].) Each subject was verbally checked for any hearing disorder. A total of 361 subjects participated in the experiment. These subjects were divided into two groups: one in whole-body photo sessions (178: 76 male subjects and 102 female subjects) and the other in face close-up sessions (183: 98 male subjects and 85 female subjects).

E. Procedure

As mentioned above, the whole experiment was divided into two sessions: W and F. Each session, in turn, consisted of two small sub-sessions: a male-speaker session and a female-speaker session. For example, a subject in W session did a male whole-body session (MW) and a female whole-body session (FW) and a subject in F session did a male face-only session (MF) and a female face-only session (FF). No one was allowed to participate in both sessions to make certain the results from W and F could be compared later without any subject interference.

The present experiment was conducted separately for each subject in the lab to ensure that every subject could see high-resolution photos and listen to the individual sound files. Each subject was seated in front of a computer monitor displaying sixteen sound files clearly labeled as *Male A* through *H* and *Female A* through *H*. Freeware *Praat* was used to display and play the sound files. Beside the monitor was located the paper with eight photographs labeled *1* through *8*.

Brief instructions on how to use the software to select and play back a desired sound file were given for each subject. The subjects listened to the sounds through a BeyerDynamics DT211 headphone set. They were permitted to listen to each sound file as many times as they wanted in any order they wanted, and asked to write down the number of photo they chose as the owner of the voice on the answer sheet (answer sheet shown in figure 5). It was explained that each photo had to be picked at least once but only once.

Name: _____ Gender _____ Photo: Whole-body

F	Voice	A	B	C	D	E	F	G	H
	Photo								
	Favorite Voice				Favorite Photo				

M	Voice	A	B	C	D	E	F	G	H
	Photo								
	Favorite Voice				Favorite Photo				

Figure 5. Answer sheet (for W session).

In addition to matching voices with photos, the subjects were asked to choose one "favorite" voice from the sound files and one "favorite" photo from the photographs. No specific definition of "favorite" was offered. This was to check whether people have a general tendency to match their favorite voice with their favorite photo.

Once a subject finished with male speakers, he/she moved to female speaker data turning to the page of the female photographs.

The time taken by subjects varied greatly from 3 minutes to 20 minutes. When asked, most subjects expressed little confidence in their answers except for a very few (less than 20).

III. Results

A. Whole-body Photo Sessions

178 subjects (76 males and 102 females) participated in the whole-body session, and the results were given in table I (for male speakers) and table II (for female speakers) as confusion matrices. Responses from male and female subjects were pooled together since there was no significant effect of subject gender. (Detailed statistical analyses will be presented later.)

In the tables, columns indicate voices and rows indicate photos. The number of responses in percentile is given in each cell. For example, when they heard the voice of *m1*, 22% of 178 subjects matched it with the photo of *m1*

Table I. Results of MW (male whole-body) session. Underlined cells represent the correct match, and shaded cells with bold numbers represent the majority match.

	Voice								
Photo	m1	m2	m3	m4	m5	m6	m7	m8	Total
m1	**22**	10	19	10	4	10	15	11	100
m2	10	14	16	9	3	10	**22**	16	100
m3	8	26	**24**	20	1	10	2	10	100
m4	19	9.6	14	12	8	23	8	6	100
m5	3	1	1	9	**65**	4	8	8	100
m6	11	**36**	12	1	2	**24**	12	2	100
m7	18	2	7	**26**	10	9	19	9	100
m8	10	2	7	13	7	10	13	**38**	100
Total	100	100	100	100	100	100	100	100	

Table II. Results of FW (female whole-body) session. Underlined cells represent the correct match, and shaded cells with bold numbers represent the majority match.

	Voice								
Photo	f1	f2	f3	f4	f5	f6	f7	f8	Total
f1	**38**	11	8	1	7	**21**	5	9	100
f2	16	**21**	19	5	1	10	14	15	100
f3	6	18	21	3	8	20	12	12	100
f4	1	9	1	**65**	0	3	11	10	100
f5	12	10	7	1	**52**	6	2	6	100
f6	20	6	**27**	0	18	20	3	8	100
f7	2	12	9	19	0	7	**35**	16	100
f8	5	**21**	8	7	3	12	18	**25**	100
Total	100	100	100	100	100	100	100	100	

(correct match), 10% with the photo of *m2*, and 8% with the photo of *m3*, etc. Underlined numbers represent the correct matches of voices with photos. Shaded cells with bold numbers indicate the majority responses regardless of whether the match is correct or not.

Remarkably, for five male (*m1*, *m3*, *m5*, *m6*, *m8*) and five female (*f1*, *f4*, *f5*, *f7*, *f8*) voices, majority picked the right photos. (For female voices, it could be counted as six correct matches instead of five since, for the voice of *f2*, two photos including that of *f2* were chosen at the same rate.) The rate of correct matches was from as low as 21% to as high as 65%. It seems reasonable to assume that people were able to identify the owner of a voice from a set of photos with a fair amount of accuracy.

However, another interesting phenomenon can be observed in the cases where the voices were not matched correctly. Even in those mismatches, there was a strong tendency for subjects to match a certain photo with a certain voice. For example, 36% of all subjects chose the photo of *m6* as the owner of the voice of *m2*, and 26% chose the photo of *m7* as the owner of the voice of *m4*. Every mismatch had a majority response with the rate of 21% or higher. This means that many people are reminded of a similar physical image from the information available in a voice alone, even though the image might not be correct in reality. It seems people had a similar principle working when they tried to match voices with photos. This may bear more important implications on the understanding of voice source than correct voice-image matches may. This confirms the results from the previous study[4] even though actual details vary a little.

B. Face-only Photo Sessions

When presented with face-only photos, the results were somewhat different as shown in table III (for male speakers) and table IV (for female speakers). The immediately noticeable difference is that the number of cases in which correct match coincides with the majority match was down for all male and female speakers. For male voices, the correct match was reduced to two (from five in W sessions) and for female voices, it was down to four. (The voice of *m8* was not counted as a correct match because the majority match of 15% is just slightly above a chance level given 8 possible choices.)

Unlike with whole-body sessions, we can see some clear difference between male and female voices. First of all, as mentioned above, more female voices were matched correctly than male voices were. Also, we can see another difference: while only one (*m5*) out of two correct matches of male voices (*m5*, *m6*) was also matched correctly with whole-body photos, all four female voices (*f4*, *f5*, *f7*, *f8*) were matched correctly with whole-body photos. And all mismatches with female whole-body photos have exactly the same pattern of mismatches with face-only photos; in both sessions, the voices of *f2*, *f3* and *f6* were matched with the photos of *f8*, *f6*, and *f1*, respectively. The same

Table III. Results of MF (male face-only) session. Underlined cells represent the correct match, and shaded cells with bold numbers represent the majority match.

	Voice								
Photo	m1	m2	m3	m4	m5	m6	m7	m8	Total
m1	16	5	11	9	27	13	9	10	100
m2	6	7	**18**	23	2	9	21	**15**	100
m3	18	**41**	16	4	1	11	1	8	100
m4	**22**	17	15	9	2	20	7	9	100
m5	6	0	4	14	**53**	2	7	**15**	100
m6	13	23	11	2	1	**23**	17	9	100
m7	10	3	11	**28**	8	8	14	19	100
m8	9	4	14	11	8	14	**25**	15	100
Total	100	100	100	100	100	100	100	100	

Table IV. Results of FF (female face-only) session. Underlined cells represent the correct match, and shaded cells with bold numbers represent the majority match.

	Voice								
Photo	f1	f2	f3	f4	f5	f6	f7	f8	Total
f1	14	22	9	2	27	**26**	8	13	100
f2	18	14	21	3	8	14	10	13	100
f3	15	17	15	4	4	11	16	17	100
f4	2	14	1	**62**	0	1	10	9	100
f5	14	4.4	15	1	**50**	10	2	3	100
f6	**25**	8.2	**27**	2	6	20	3	8	100
f7	8	19	8	9	1	8	**30**	17	100
f8	5	**20**	3	16	4	10	21	**20**	100
Total	100	100	100	100	100	100	100	100	

is true only for the voices of *m2* and *m4*. For some reason, female voices seem to be more consistently matched.

Majority matching rate is higher in W (average of 35%) than in F (average of 29%), and it is higher with female speakers (average of 35%) than with male speakers (average of 29%).

C. Statistical Analysis: Generalized Estimated Equation

To verify the validity of the observations above, two statistical analyses were conducted. First, one-way Chi-square values were evaluated for each cell in tables I through IV to test the goodness of fit. The results are not presented here because every cell with the majority vote is well above significant level ($P<.005$). (It turns out that any cell with less than 4% or more than 20% is statistically significant.)

There are three independent variables in this experiment: speaker gender, sessions (W/F), and subject gender. Since the responses are discrete, as well as independent variables, the generalized estimated equation method is used to check intra-variable effects. Two separate calculations are made: one for correct responses, and the other for majority responses. In each case, only the relevant (that is, correct or majority) response was coded as 1 and the rest were coded as 0. The results are shown in table V. As we can see in the table, all variables except subject gender played important roles. Speaker gender effect is particularly strong

Table V. Statistical analysis by generalized estimated equation method.

variable	Pr <		
	speaker gender	session	subject gender
correct response	0.0001	0.0001	0.2580
majority response	0.0570	0.0005	0.1404

when we consider the correct response case. We can safely say that these statistical analyses support our previous impression: that is, speaker gender and the type of photo presented play very important roles in matching voices with photos.

D. "Favorite" Voice vs. "Favorite" Photo

As previously mentioned, subjects were asked to identify their favorite voice as well as their favorite photo. To verify whether people generally match their favorite voice with their favorite photo, a response from each subject was individually examined in the following manner: If a subject picked *A* as his favorite voice and photo number *2* as his favorite photo, and if he had chosen photo number *2* as the owner of the voice *A*, then this is a 'positive' case in which a favorite voice coincides with a favorite photo. Any answer other than *2* was considered as a 'negative' case. If there are more positive cases, we might have to conclude that the matches or mismatches shown above were based on subjective judgment of matching favorite voices with favorite photos.

The results, however, indicate otherwise. Positive cases are less than a third (28-32%) in most instances except for female face-only, which was 46%. This clearly shows that subjects did not just automatically match their favorite voice with their favorite photo.

Close examination of this data from different perspective reveals a very interesting point. Instead of counting positive vs. negative cases, a simple frequency count was conducted on the favorite voices and favorite photos. The results are presented in figure 6 (favorite voice) and figure 7 (favorite photo). In the figures, the percentage of a certain speaker chosen as favorite voice or favorite photo is plotted, and empty columns represent the results from W sessions while filled columns represent the results from F sessions. For example, in figure 6, the voice of *f1* was chosen as the favorite voice by 23% (41 out of 178) of subjects in W sessions and by 14% in F sessions.

Comparing figures 6 and 7, we can immediately notice that subjects have much less diverse opinions on which voice they like most than on which photo they like most. Favorite photos change dramatically depending on the types of photos. For example, the photo of *f4* was a clear winner in F sessions, but the votes for the photo of the same person in W sessions dropped dramatically. This trend can be observed in male data as well even though

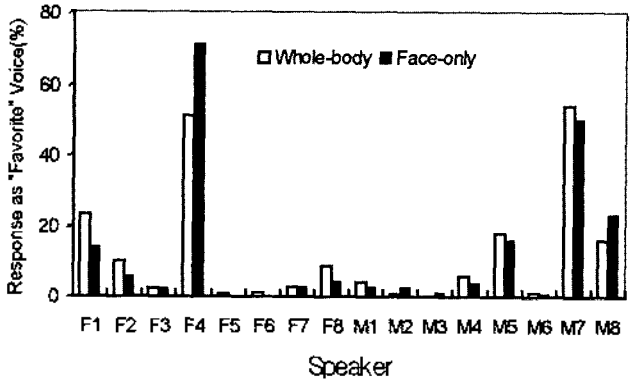

Figure 6. Percentage of a certain speaker chosen as the favorite voice (Empty columns represent W sessions and filled columns represent F sessions.)

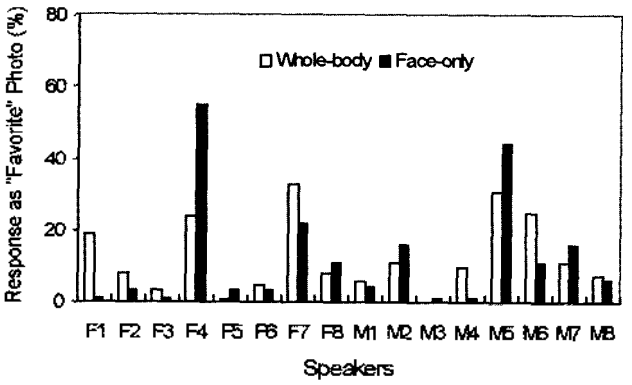

Figure 7. Percentage of a certain speaker chosen as the favorite photo (Empty columns represent W sessions and filled columns represent F sessions.)

with a little lesser degree. There were many contestants fighting for the "favorite photo" title in both males and females. However, this was not the case with favorite voice. Two groups of subjects from different sessions chose the same voice as their favorite voice. The votes for the favorite voice seemed to be almost unanimous and there was no real competition: the voices of *f4* and *m7* were favored by majority of subjects.

IV. Discussion

A. Summary

To summarize, we have observed the following:

(1) Majority of the subjects in this experiment matched 5 male voices and 5 females voices correctly with their corresponding whole-body photos.

(2) Even in the cases in which voices were matched incorrectly, there was a strong tendency for majority of the subjects to match a certain photo with a certain voice.

(3) The results were very different depending on what type of photo was presented: both the accuracy of correct matches and the rate of majority matches were higher with whole-body photos than with face-only photos.

(4) In both sessions, the results for female speakers showed both higher correct matching rates and higher majority matching rates.

(5) The subjects did not necessarily match their favorite voices with their favorite photos.

(6) The subjects were much more like to agree on their favorite voice than on their favorite photo.

B. Discussion

From (1) and (2) above, we may conclude that, hearing a voice, different people make up, or expect a similar appearance from the voice alone, at least in an environment in which they are forced to choose one among many possibilities. This indicates that a voice does carry certain information about the physical characteristics of the person, and not only are listeners capable of capturing this information but they also share a common mechanism to translate the voice into an image.

This result was quite a surprise to many people. Prior to the experiment, the author informally asked people about their experiences. Most said that they were more often surprised to meet the person, because their expectations were wrong. But according to this study, that should not be the case. It might be that the people the author interviewed simply remembered more vividly the instances when they were surprised than the instances when they were not.

The different results from whole-body photo sessions and face-only photo sessions summarized in (3) above are very revealing. The main difference between whole-body photos and face-only photos is that face-only photos do not carry the information on physical characteristics such as size and height as much as whole-body photos do. Therefore, we can assume that subjects were able to perform better when they were given more information on physical characteristics.

This may have an interesting implication on understanding the richness of the vocal source. As mentioned in *I. Theoretical consideration*, vocal folds are considered to be a source of sound and vocal tract is considered to be a filter in the source-filter theory of speech production. And naturally the filter characteristics are closely related with physical characteristics. Therefore this result seems to be in line with what the theory would have predicted.

However, is this perception the result of deducing and processing filter characteristics of the speaker, or the result of processing only the source? In other words, can the source have enough information about physical characteristics of the speaker all by itself? Or at least, could the source be the primary provider of the relevant information? Since vocal folds are a part of a body, it is only natural to assume that they should reflect the owner's physical condition somehow. As with the current study, we cannot answer this question because the data we saw were the results from the perception experiment based on natural utterances that were already the mixture of source and filter characteristics. This question may be pursued with

an experimental design with a synthesizer in which filter characteristics can be remain the same while only the source characteristics are changing.

The big difference between male voices and female voices summarized in (4) poses a question which cannot be answered easily either. Why is it easier to match female voices?

It is interesting to note that the male speakers were much more diverse in their physical differences than the female speakers. The fact that the number of correctly matched cases dropped from 5 in W session to 2 in F sessions strongly suggests that physical characteristics were needed to have correct matches. On the other hand, the results from female speakers seem to suggest something else than physical characteristics. They showed much less differences in their physique than male speakers. Still matching rate (both correct and majority) was higher for females both in W and in F sessions.

Age might be responsible for this result. Female speakers were more diverse in their age than males, and thus they were more easily identified. However, it is not certain whether the subjects were able to judge the speakers' ages either by listening voices or by looking at the photos. This could be simply an idiosyncratic phenomenon to the current experiment and not a general tendency.

As an alternative explanation to the observations (3) and (4), several people mentioned the 'fashion'. According to them, fashion can make a big impression and thus influence the judgment of the subjects. (Being a male completely ignorant of fashion, the author did not take fashion into account when the speakers' photos were taken. It was also not possible for the author to tell the speakers what to wear for the photos.) This could explain the observation (4), because female speakers were more diverse and expressionistic in their clothes then were males. And it might also partly explain (3), since it is easier to see the clothes in whole-body photos. But that explanation immediately raises another question: what kind of fashion is related to what kind of voice? Nonetheless, this should be considered as another factor which might influence the judgment.

Observation (5) contradicted the author's informal interviews again. Many people believed they would match a good voice with a favorable appearance. This seems to be another unfounded myth. There must be something concrete and substantial for such a large number of subjects to show a different trend, as summarized in (5).

Another important finding was summarized in (6). Why is it easier to agree on a favorite voice than to agree on a favorite photo? The answer may be a long way off. However, once we take this observation as an objective finding, it suggests something very interesting along with findings (1) and (2), especially for speech technology. Preference of appearance seems to be influenced by many factors such as the photo setting and/or the clothes. However the favorite voice remained constant by more than 50% of the subjects, regardless of conditions of presentation. This implies that it would be easier to synthesize a voice favored by the majority than to create an image favored by the majority.

C. Comparing the Present Study with Moon (1999)[4]

The present study confirms all the findings in the previous study. Even though almost four times more number of subjects participated in the current experiment, basic findings are still the same.

However, there do exist some differences if we look at the data very closely. For example, the majority match for *m1* was the picture of *m3* in the earlier study, but it is now *m1*, which is also the correct match. For each condition, the voices which have the same majority matches as Moon (1999) are as follows: for MW, 5 voices (m3, m4, m5, m6, and m8), for FW, 6 voices (f1, f4, f5, f6, f7, and f8), for MF 6 voices (m1, m2, m3, m5, m6, and m8), and for FF, 5 voices (f1, f4, f5, f6, and f7). These differences might be attributed to the different experimental settings in which much higher resolution photographs were presented for each individual subject. It is still surprising rather than disappointing that, in different settings, such a many number of people, who were two completely exclusive populations, agreed on who must have produced a certain voice. It confirms the major finding that there definitely IS a certain relationship between voices and the images triggered by the voices.

V. Conclusions and Suggestions

The present study is limited in the sense that it did not control several factors such as age and 'fashion', which might have influenced the subjects' judgment. Future research on the subject should be more carefully controlled.

This limitation notwithstanding, this study clearly shows that a voice causes listeners to create a certain image, and that the image is very similar among many people. This suggests that the psychological effect of voice is not just an illusion, but the result of a mechanism based on the physical and concrete properties of the voice.

If we accept the findings of this study, this topic of voice-image matching deserves further research. With the advancement of technologies, we are seeing, or hearing, more and more applications of speech synthesis. The present finding suggests that we might not have to put up with a certain stereotypical voice for computer-generated speech. If we can find more substantial correspondences between voice qualities and the images matched with those qualities (and thus their psychological effects in some broad sense), we can use appropriate voices for different situations. This may be a wishful thinking today. However, it is established in this study that the relationship between a voice and the image conjured by the voice is not random, but rather very closely related in a manner which many people share. Therefore it should be possible, with a systematic approach toward the subject, to map this relationship in a more detailed way. For this purpose, much more narrowly focused research should be undertaken with systematic control of variables.

This line of research will contribute not only toward advancement of speech technology but also toward the enhancement of the phonetic and psychological understanding of the human vocal source.

Acknowledgment

This work was supported by the Korea Research Foundation Grant (KRF-1998-001-A00006) and in part by Ajou University Grant. The author is also indebted to Nate Marti in Statistical Consulting Service at the University of Texas at Austin who patiently helped with statistical analyses. (Needless to say, all the faults, if there are any, are mine.)

References

1. J. Sundberg, *The Science of Singing Voice*. Northern Illinoise University Press, Dekelb, 1987.
2. L. Leinonen, and T. Hiltunen, "Expression of emotional motivational connotations with a one-word utterance," *Journal of the Acoustical Society of America*, **102** (3), pp. 1853-1863, 1997.
3. S-J. Moon, "Voice and image: a pilot study," *Malsori (Phonetics)*, **35-36**, pp. 37-48, 1998.
4. S-J. Moon, "Voice and image: A perception experiment," *Journal of the Acoustical Society of Korea*, **18.8**, pp. 66-74, 1999.

[Profile]

• Seung-Jae Moon

The Journal of the Acoustical Society of Korea, Vol. 18, No. 8.