말소리와 음성과학 제 1 권 제 2 호 (2009)
pp. 111~118

목소리를 듣고 감지하는 인상에 대한 연구: 미국인화자와 미국인청자

The Relationship Between Voice and the Image Triggered by the Voice: American Speakers and American Listeners

문 승 재 [1)]
Moon, Seung-Jae

ABSTRACT

The present study aims at investigating the relationship between voices and the physical images triggered by the voices. It is the final part of a four-part series and the results reported in the present study are limited to those of American speakers and American listeners. Combined with the results from previous studies (Moon, 2000; Moon, 2002; Tak, 2005), the results suggest that (1) there is a very strong, much higher than chance-level relationship between voices and the pictures chosen for the voices by the perception experiment subjects; (2) the more physical characteristics that are given, the better the chance for correctly matching voices with pictures; and (3) culture (in the present, language environment) seems to play a role in conjuring up the mental images from voices.

Keywords: voice, voice quality, speaker recognition; emotion; emotion technology

1. 서 론

1.1 목소리와 청각이외의 다른 요소와의 관계

우리는 사람의 목소리를 듣고 그 사람에 대하여 많은 것을 알 수 있거나 혹은 짐작하게 된다. 그 사람이 그 전부터 알던 사람이라면 말소리만 듣고도 그 사람이 누구인지 금방 알 수 있으며, 모르는 사람의 경우라도 그 말하는 사람의 나이, 감정상태 등에 대해서 짐작할 수 있다. 본 연구는 목소리가 전달하는 이처럼 다양한 정보 중에서도 특별히 말하는 사람의 외모와 관계되는 정보와의 관계를 살펴보고자 한다.

목소리에 대한 연구는 다방면에서 많았지만, 이처럼 목소리와 그로 인하여 연상되는 외모와의 관련성을 연구한 예는 아직 필자외에는 없는 것 같다.

목소리는 단순히 언어적인 정보를 전하는 도구로 쓰일 뿐 아니라, 매우 풍부한 부(副)언어적인 요소를 전달한다 (최양규와 신명선, 2004). 실제 생활에서는 이러한 부언어적 요소가 언어적 정보보다 훨씬 더 많은 정보를 전달하게 된다. 단적인 예로, 일반적으로 말로 전달되는 정보는 7%인 반면, 목소리로 38%, 표정으로 55%의 정보가 전달된다고 Mehrabian(1973)은 보고하였다 (탁지현, 2005 에서 재인용).

이처럼 목소리가 언어활동에서 차지하는 역할에 대한 연구는 상당히 광범위하게 다방면에서 이루어져왔다. 그 중에서도 본 연구와 관점이 비슷한 청각적인 소리와 시각적인 정보에 관한 연구를 살펴보면, Kuhl & Meltzoff (1982, 1984)는 생후 5 개월 미만의 아기들에게 말소리는 시각적인 정보와 연결지을 수 있는 청각적인 정보를 제공해준다는 것을 밝혔으며, Green et al. (1991)은 남자의 목소리를 여자의 얼굴과 함께 제시한 McGurk 효과 실험에서, 비록 소리와 화면의 정보가 상이한 경우에도 McGurk 효과는 소리와 화면의 성별이 같을 때와 마찬가지로 존재하였음을 보여주고, 이것은 말하는 사람의 목소리에 대한 정보가

1) 아주대학교 moon@ajou.ac.kr

접수일자: 2009 년 4 월 30 일
수정일자: 2009 년 6 월 3 일
게재결정: 2009 년 6 월 20 일

청각과 시각 정보가 합쳐지기 전 단계에서 이미 추출되고 이용됨을 암시한다고 하였다.

목소리가 전달하는 그 목소리의 주인공에 대한 정보 중 일반적으로 매우 정확한 것은 나이에 대한 것으로 사람들은 목소리를 근거로 말하는 사람의 나이를 매우 정확하게 짐작해낸다(Linville, 2000). 사람들은 목소리를 듣고 말하는 사람의 감정상태 역시 상세한 상태로 알아낼 수 있다(Leinonen& Hiltunen, 1997; Murray & Arnott, 1993; Scherer et al., 1991).

공학에서는 목소리와 시각적정보와의 관계를 이용하여 좀 더 자연스러운 시각적 컴퓨음성합성기를 개발하기 위한 연구를 진행하여왔다. Cohen, Beskow & Massaro(1998)는 컴퓨터를 이용하여 말하는 사람의 얼굴을 개발하는 과정에서 구체적인 조음적 특징, 특히 구개, 이, 혀의 모습을 투명한 얼굴을 통하여 실감나게 현실적으로 구현함으로써, 청각장애아들의 언어교육에 도움이 될 수 있음을 보여주었다.

여러 종류의 연구 중 본 연구와 가장 관계가 깊은 것은 아마도 화자인식분야일 것이다. 화자 인식에 대해서는 매우 많은 연구가 있어왔다 (Fellowes, Remez, & Rubin, 1997; Remez, Fellowes, & Rubin, 1997; Remez et al., 2001 등). 자연스러운 발화에서 언어적요소를 제거한 sinewave speech 를 이용한 일련의 실험에서 Remez et al. (1997)은, sinewave speech 만으로도 이미 알던 사람의 경우는 화자를 정확히 인식할 수 있다는 것을 관찰하였다. 이러한 결과로부터 연구자들은, 화자를 인식하게 하는 중요한 정보는 언어적요소가 아닌 시간에 따라 변하는 음성학적 요소(time-varying phonetic information)에 내재되어 있다고 주장하였다.

1.2 목소리와 외모

위에서 살펴본 연구들은 모두 목소리와 청각 이외의 다른 요소들의 관계를 연구하였다. 그러나 본 연구처럼 목소리와 그 목소리로 인하여 듣는 사람이 그려내는 외모와의 관계를 연구한 예는 없었다.

우리는 아는 사람의 목소리를 들으면 그 즉시 그 사람의 모습을 떠올리게 된다. 그러나 모르는 사람의 목소리를 들었을 때는 어떤가? 모르는 사람의 목소리를 처음으로 들었을 때, 최소한 본 연구자는 은연중에 말하는 사람의 외모나 성격에 대한 인상을 갖게 된다. 그러다가 그것이 나만의 현상인지 아니면 남도 그러는지, 그리고 그럴 때에 떠오르는 인상이 과연 정확성이 있는 것인지에 대한 의구심이 생겼다.

이러한 의구심을 풀고자 하는 것이 본 연구의 출발점이었다. 그러나 사람들이 어떤 한 목소리를 듣고 어떤 모습을 연상하는지를 확인할 현실적인 방법이 없었다. 사람들에게 목소리를 듣고 떠오르는 모습을 기술하라고 할 수도 없고, 그려보라고 할 수도 없는 일이기 때문이다. 혹 수사기관에서 사용하는 몽타쥬 그리는 방법을 쓰면 가능한 일인지 모르겠으나, 많은 사람들에게 그런 결과를 얻는다는 것은 현실적으로 불가능한 일이었다. 다른 더 좋은 대안이 없기 때문에 고육지책으로 생각해낸 것이, 사람들을 직접 녹음하고 그 주인공들의 사진을 찍은 후, 그 사람들을 모르는 사람들에게 목소리를 들려주고 사진과 연결하도록 하는 것이었다. 목소리를 듣고 실제 있는 사진 중에서 그 주인공을 고르도록 한 것이 그 목소리를 듣고 형성되는 "인상"과 얼마나 깊은 관계가 있을 것인가에 대한 이론의 여지가 있을 수 있겠다. 그러나 본 연구에서는 "인상"을 *image* 의 개념으로 보았고, 현재 그것을 확인하는데 가장 근접할 수 있는 방법이 이 방법밖에 없다고 판단하여 연구를 진행하였다.

이러한 인상에 대한 정확성이 얼마나 되는지, 그리고 만일 그것이 어느 정도의 정확성을 갖는다면 어떻게 그러한 '목소리-외모'간의 연결이 가능한지를 규명하기 위하여, 다음과 같은 일련의 연구가 진행되어 왔다: 말하는 사람과 그 목소리를 듣는 사람에 따라, 한국인이 말한 소리를 한국인에게 들려준 연구(Moon, 2002), 같은 한국인의 자료를 한국어를 모르는 미국인에게 들려준 연구(Moon, 2000), 미국인이 말한 소리를 한국인에게 들려준 연구(탁지현, 2005)가 있었다. 본 연구에서는 탁지현(2005)에서 사용된 미국인의 소리를 미국인에게 들려주고 목소리와 사진을 짝짓도록 한 결과를 보고함으로써 한국인/미국인화자 – 한국인/미국인청자 간에 가능한 네 가지 경우에 대한 결과보고를 완결짓고자 한다.

2. 인지실험

본 연구에서 쓰인 자료는 탁지현(2005)의 연구에서 쓰인 자료와 동일한 자료이다.

2.1 화자

화자는 수원 근처에 소재하는 어학원에서 강사로 일하고 있는 25-33 세의 미국인 남자 8 명과 25-34 세의 미국인 여자 8 명이었다. 목소리에서 쉽게 짐작할 수 있는 나이의 차이를 최소화하기 위해 노력했다.

2.2 말자료

말자료로는 위의 화자들이 영어권내에서 언어치료 진단 등에 흔히 쓰이는 "The Rainbow Passage"(Kent, 1994)를 읽은

것을 사용하였다. 화자들이 이를 읽는데 걸린 평균시간은 약 2분 30초 정도였으며 이 중 남녀화자별로 각각 다른 부분 약 15 초 분량을 선택하여 인지실험에 사용하였다. 이는 남녀 화자 모두를 들어야 하는 인지실험 참여자들을 고려하여 덜 지루하도록 하기 위함이었다.

2.3 사진자료

사진자료는 녹음 직후 촬영하였다. 사진은 200 만 화소 디지털 사진기(Sony Cybershot U-20)를 사용하였으며, 키 등의 신체조건을 보존하기 위하여 2.1m 높이의 출입문 앞에서 촬영하였다.

목소리와 연결짓도록 제시된 사진은, 키, 체격 등 신체적 특징이 잘 드러나는 전신사진과, 신체적 특징이 잘 드러나지 않는 얼굴사진의 두 종류였다. 전신사진은 각자의 사진을 촬영한 후, 별도의 이미지 편집 프로그램을 이용하여 키를 근거로 신체의 상대적 크기를 유지한 상태에서 함께 모았으며, 얼굴사진은 최종사진의 크기가 비슷하도록 모았다.

인지실험에 참여한 누구도 전신사진과 얼굴사진의 두 세션을 동시에 참여하지 않도록 하였다.

인지실험에 이용된 전신사진과 얼굴사진은 다음과 같다. (실제 실험에서는 이 사진들을 각각 가로 21.59cm, 세로 27.94cm 의 'letter-size' 종이 한 면에 꽉 차도록 천연색으로 출력하여 사용하였다.)

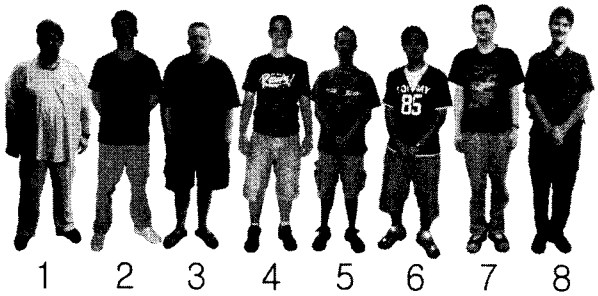

그림 1. 남자전신사진
Figure 1. Male whole-body picture

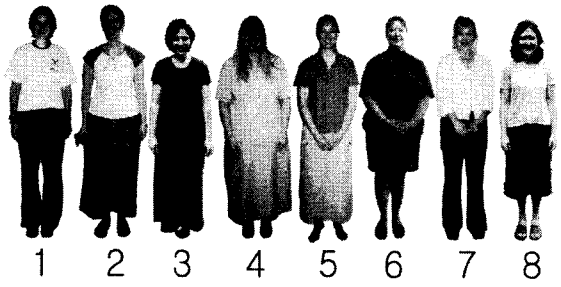

그림 2. 여자전신사진
Figure 2. Female whole-body picture

그림 3. 남자얼굴사진
Figure 3. Male face close-up picture

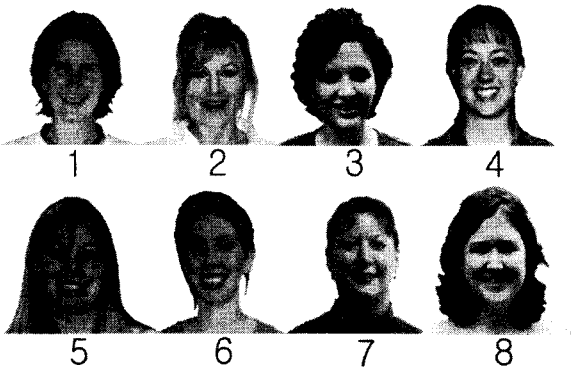

그림 4. 여자얼굴사진
Figure 4. Female face close-up picture

2.4 인지실험

인지실험 참여자는 University of Massachusetts at Amherst 에 재학중인 학부생 73 명으로서, 이들은 언어학개론과목을 듣는 과정의 일부로서 이 실험에 참여하였다. 인지실험참여자는 스스로 판단하기에 청력에 이상이 없는, 영어를 모국어로 하는 사람으로 한정하였으며, 37 명은 전신사진세션, 36 명은 얼굴사진세션에 투입되었다.

인지실험은 University of Massachusetts at Amherst 의 Phonetics Lab 에 설치되어있는 인지실험실에서 한번에 최대 4 명까지를 대상으로 실시되었다. 각자의 모니터에 사진과 소리파일을 나타내는 스피커 아이콘이 나타나서, 참여자는 원하는 소리파일을 클릭해서 들으면서 사진을 볼 수 있도록 하였다 (실제 실험에 사용된 화면의 한 예인 <그림 5> 참조).

참여자들은 주어진 사진을 보고 주어진 소리파일을 선택해 들으면서 각각의 목소리에 대해 그 주인공이라고 생각되는 사진의 번호를 적도록 지시를 받았는데, 소리파일을 듣는 횟수나 순서, 전체 실험을 마치는데 허용되는 시간 등의 제한은 받지 않았다. 그러나 각 실험참여자에게 한 목소리에 대해서는 한 사진 만을 고를 수 있음을 주지시켰다.

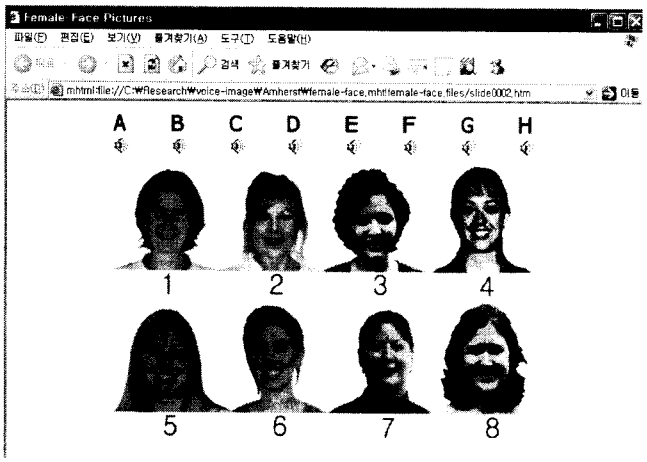

그림 5. 여자얼굴세션 실험화면
Figure 5. A screen capture of female face close-up session

목소리와 얼굴을 연결하는 이외에도, 각 참여자는 가장 좋은 목소리("favorite voice")와 가장 좋은 사진("favorite picture")을 하나씩 고르도록 하였는데, 이 때 "가장 좋은"의 정의는 따로 주지 않고 각자 직관적으로 판단하도록 하였다. 이것은 좋은 목소리와 좋은 모습을 연결짓는 경향이 있는지를 보고자 함이었다.

3. 결과 및 논의

3.1 판단을 위한 기준

본 연구의 결과를 기술하기에 앞서 본 연구의 결과를 어떻게 해석할 것인가에 하는 문제에 대해 먼저 기술하고자 한다. 본 연구의 실험의 결과는 인지실험 참여자들이 여덟 개의 목소리를 듣고, 주어진 여덟 사람의 사진 중에서 하나를 고른 양상으로 나타나는데, 과연 세션별 결과를 어떻게 해석하며 다른 세션의 결과를 어떻게 서로 비교할 것인가 하는 것이다.

우선, 본 인지실험의 판단결과가 계속적인 값(예를 들면 모음의 길이)이 아니라 별개의 선택(사진)이었기 때문에, 각 세션별로 여덟 개의 선택 중 하나를 한 것이 과연 유의미한 선택인지를 확인하기 위해서 카이스퀘어 검증을 실시하였다. 그 결과 모든 세션에서 각 목소리와 각 사진의 연결분포는 유의미한 차이를 보여주었다. 각 세션별 X^2 값과 유의수준은 다음의 <표 1>과 같다.

표 1. 세션별 Chi-Square 검증 결과
Table 2. Chi-Square test results for each session

세션	남성전신	여성전신	남성얼굴	여성얼굴
X^2	263.135	174.919	325.333	148.444
유의수준	.000	.000	.000	.000

이로부터 인지실험 결과에서 나타나는 선택의 분포는 통계적으로 유의한 차이임을 알 수 있었다. 즉, 모든 세션에서 실험참가자들이 목소리에 대해 차별적으로 추정하는 인상(혹은 사진)이 존재함을 확인할 수 있었다.

그러나 이 차원을 넘어서서, 과연 전신세션에서 보인 선택결과와 얼굴세션에서 보인 선택결과의 차이가 통계적으로 유의미한 것인지를 밝혀낼 수 있는 통계적 방법은 찾지 못하였다. 이를 위하여 몇몇 기관의 통계전문 상담센터를 방문했으나 현재의 실험과 같은 설계하에서는 이러한 질문에 답을 줄 수 있는 방법이 없다는 답을 들었다. 따라서 개별적인 세션별 결과를 벗어나는 차원의 비교나 기술에 대해서는, 특별한 통계적 자료를 갖고 있지 못한 실정이다.

그러나 이러한 제한점이 본 연구의 결과의 의미를 약화시키지는 않는다고 본다. 본 연구의 결과에 보고되는 내용은, 각 세션별 응답의 분포가 통계적으로 유의하다는 카이스퀘어 검증에 바탕을 두고, 그 중에서도 가장 큰 값들을 중심으로 논의할 것이기 때문이다.

3.2 전신사진세션

남녀 화자에 대한 전신사진세션의 실험결과는 각각 <표 2>, <표 3>과 같다.

이 표는 각 목소리(MA, MB, ... 등)를 듣고 참여자들이 고른 사진(1,2, ... 등)의 빈도수를 퍼센트로 표시한 것이다. 예를 들면 MA 의 목소리에 대해 22%의 참여자는 사진 1 을 연결한 반면 3%의 참여자는 사진 2 또는 사진 4 를 연결하였다.

표 2. 인지실험결과: 남자화자 전신사진세션
Table 2. Results from male-speakers whole-body picture session

	목소리								
사진	MA	MB	MC	MD	ME	MF	MG	MH	계
1	22	27	3	8	3	16	16	5	100
2	3	5	22	14	11	19	5	22	100
3	11	16	3	19	14	19	0	19	100
4	3	3	32	3	19	8	3	30	100
5	16	14	3	16	32	14	3	3	100
6	0	0	22	3	3	0	70	3	100
7	8	14	16	19	11	11	3	19	100
8	38	22	0	19	8	14	0	0	100
계	100	100	100	100	100	100	100	100	

표 3. 인지실험결과: 여자화자 전신사진세션
Table 3. Results from female-speakers whole-body picture session

	목소리								
사진	FA	FB	FC	FD	FE	FF	FG	FH	계
1	3	11	24	8	38	5	3	8	100
2	3	30	19	19	3	11	3	14	100
3	14	8	11	8	5	19	8	27	100
4	8	24	14	14	14	5	14	8	100

5	11	11	8	19	5	14	14	19	100
6	0	5	8	22	27	3	24	11	100
7	11	3	16	3	5	24	30	8	100
8	51	8	0	8	3	19	5	5	100
계	100	100	100	100	100	100	100	100	

표에서 어둡게 칠해진 부분은 각 목소리에 대해 가장 많은 참여자가 연결지은 사진을 가리키며 (이를 앞으로 '최다선택'이라고 함), 응답률에 밑줄이 쳐진 것은 그 목소리의 실제 주인공을 가리킨다 (이를 앞으로 '정답'이라고 함). 예를 들면, 목소리 MA 의 실제 주인공은 사진 8 번이며, 이 경우에는 가장 많은 참여자(38%)가 사진 8 번을 그 목소리와 연결하였음을 보여준다.

정답과 최다선택이 일치하는 경우를 보면 남자화자의 경우 다섯 명, 여자화자의 경우 네 명이었다. (남자화자 MD 의 경우는, 최다선택이 사진 3, 7, 8 에 대해 동일하게 나왔으나, 그 중 정답인 경우도 있기 때문에 정답과 최다선택이 일치하는 경우로 보았다.)

이 결과를 선행연구들과 비교해보자. 같은 영어자료를 한국인에게 들려주었던 탁지현(2005)의 경우는 정답과 최다선택이 일치하는 경우가 남자화자 여자화자 각각 세 명씩이었다. 한국어 자료를 각각 한국인과 미국인에게 들려주었던 Moon(2002)과 Moon(2000)을 비교한 결과와 같은 추세를 보여준다. 한국어 자료를 한국인이 들었을 때에는 정답과 최다선택이 일치하는 경우가 남녀화자 각각 다섯 명씩이었던 반면, 같은 자료를 미국인에게 들려주었을 때에는 일치한 경우가 남자화자는 두 명, 여자화자는 한 명이었다. 이 두 결과를 종합해볼 때, 목소리를 듣고 외모를 판단하는 과정에 문화적인 차이, 즉 이 경우는 언어에 관계되는 차이가 존재하여, 자기나라의 말을 들은 사람들이 외국인보다 더 정확한 결과를 도출해내었다는 것을 보여주고 있다고 할 수 있을 것이다.

3.3 얼굴사진세션

36 명의 참여자가 얼굴사진을 보고 목소리와 연결을 지은 결과는 다음의 <표 4>, <표 5>와 같다.

표 4. 인지실험결과: 남자화자 얼굴사진세션

Table 4. Results from male-speakers face close-up picture session

	목소리								
사진	MA	MB	MC	MD	ME	MF	MG	MH	계
1	14	28	0	19	8	17	0	14	100
2	36	8	3	6	14	17	11	6	100
3	22	11	6	22	25	11	0	3	100
4	6	0	25	17	25	6	6	17	100
5	6	8	33	0	8	3	3	39	100
6	6	0	17	0	0	0	78	0	100
7	8	39	3	17	3	28	3	0	100
8	3	6	14	19	17	19	0	22	100
계	100	100	100	100	100	100	100	100	

표 5. 인지실험결과: 여자화자 얼굴사진세션

Table 5. Results from female-speakers face close-up picture session

	목소리								
사진	FA	FB	FC	FD	FE	FF	FG	FH	계
1	8	14	11	6	39	3	6	14	100
2	36	3	19	3	0	14	22	3	100
3	8	11	6	6	8	17	19	25	100
4	17	19	8	19	6	14	3	14	100
5	17	11	14	6	6	19	19	8	100
6	6	28	17	19	6	19	3	3	100
7	3	6	6	28	25	3	28	3	100
8	6	8	19	14	11	11	0	31	100
계	100	100	100	100	100	100	100	100	

얼굴사진세션에서 정답과 최다선택이 일치하는 경우를 보면 남자화자 세 명, 여자화자 세 명이었다. 이 결과를 선행연구들과 비교하여 보자. 같은 영어자료를 한국인에게 들려준 탁지현(2005)에서는 정답과 최다선택이 일치하는 경우가 남자화자 한 명, 여자화자 두 명이었다. 우리말 자료를 한국인에게 들려준 Moon(2002)의 경우는 남자화자의 경우 두 명, 여자화자의 경우 세 명이었고, 같은 우리말 자료를 미국인에게 들려준 Moon(2000)의 경우, 남자화자는 두 명, 여자화자는 한 명이었다.

전신과 얼굴사진세션 모두에서 압도적으로 목소리와 사진이 제대로 연결된 경우는 목소리 MG 로서 70% 이상의 응답자가 올바른 사진을 선택하여 다른 어떤 화자들과도 다른 매우 높은 정확성을 보였다. 실험 후에 확인한 결과, 미국인들은 목소리 MG 에 대하여 남미계의 강한 액센트가 있기 때문에 그러한 모습의 사람을 골랐다고 답변하였다. 그런데 이것은 같은 자료를 이용하여 한국인들에게 소리를 들려준 탁지현(2005)의 연구결과와는 매우 대조적이었다. 탁지현(2005) 에서는 같은 MG 의 목소리에 대하여 최다선택이 전신사진세션에서는 6 번(26%), 얼굴사진세션에서는 3 번(24%)이었다. 즉 한국인들은, 미국인들에게 명백히 드러났던 소위 '액센트'의 차이를 그리 크게 느끼지 못했음을 보여주었다. 이것은 우리나라 사람들이 남미식의 영어를 자주 접해보지 못했던 문화적인 차이의 예라고 하겠다.

3.4 연구별, 세션별 최다선택과 정답 일치 비교

본 연구에서 최다선택과 정답이 일치하는 경우의 수를 선행연구들과 비교하여 표로 제시하면 다음의 <표 6>과 같다.

표 6. 전신사진세션 선행연구와의 최다선택/정답 일치 비교
Table 6. Comparison of identical majority-/correct-matches in 4 studies

연구		본 연구	탁지현 (2005)	Moon (2002)	Moon (2000)
화자-청자		미국-미국	미국-한국	한국-한국	한국-미국
전신 세션	남자화자	5	3	5	2
	여자화자	4	3	5	1
얼굴 세션	남자화자	3	1	2	2
	여자화자	3	2	3	1

<표 6>을 보면 다음과 같은 공통적인 추세를 볼 수 있다.

(1) Moon(2000)을 제외한 세 연구에서는 전신세션에서 얼굴세션보다 최다선택과 정답이 일치하는 경우가 더 많다. 이것은 아마도 키나 체격 등 신체적 특징이 나타나는 경우에 응답자들이 더 정확하게 목소리의 주인공을 짐작해낸 것이 아닌가 추측할 수 있다. 물론 체격뿐 아니라 옷차림 등 다른 요소의 영향이 있었음도 배제할 수는 없으나, 그러한 여러 가지 요소 중 체격이 영향을 미쳤음을 부정할 수는 없을 것이다.

(2) 모든 연구에서 공통되게 화자와 청자의 언어권이 같은 경우가 다른 경우보다 최다선택과 정답이 일치하는 경우가 많았다. 이는 앞서 언급했던 것처럼, 언어환경에 따른 차이가 역할을 하고 있음을 볼 수 있다. 이러한 현상은 2 개국어를 모국어로 하는 사람이 말을 할 경우 듣는 사람은 자기가 익숙한 언어로 할 때 더 잘 알아듣는다는 Goggin et al. (1991)의 연구와 일관된 현상이었다.

3.5 가장 좋은 목소리와 가장 좋은 사진

가장 좋은 목소리와 가장 좋은 인상을 고르게 한 원래의 의도는, 혹시 사람들에게 가장 마음에 드는 목소리와 가장 마음에 드는 사진을 무의식적으로 연결짓는 경향이 있는가를 확인하기 위함이었다. 이를 위하여 각 참여자의 응답에서 가장 좋은 목소리로 꼽힌 목소리와 그 목소리를 듣고 연결한 사진이 가장 좋은 사진으로 꼽힌 사진과 일치하는지를 일일이 확인하였다. 예를 들면 한 참여자가 가장 좋은 목소리로 A 목소리를 꼽고, 그 A 목소리를 연결지은 사진이 1 번일 때, 그 참여자가 가장 좋은 사진으로 꼽은 사진이 1 번이라면 그 때는 이 참여자의 경우 가장 좋은 목소리와 가장 좋은 사진을 연관지었다고 보았고, 그렇지 않은 경우는 둘 사이에 아무런 연관성이 없다고 보았다.

가장 좋은 목소리와 가장 좋은 인상의 일치도를 확인한 결과는 다음의 <표 7>과 같이 요약할 수 있다.

표 7. 가장 좋은 목소리와 가장 좋은 인상의 일치도
Table 7. % of favorite voice being matched with favorite picture

	전신사진세션		얼굴사진세션	
성별	남자	여자	남자	여자
일치	30%	16%	33%	37%
불일치	70%	84%	67%	63%
계	100%	100%	100%	100%

<표 7>에서 보듯이 인지실험 참여자들은 단순히 마음에 든다는 것 외에 무엇인가 다른 이유로 목소리와 사진을 연결하였음을 알 수 있었다.

이러한 결과는 역시 선행연구(Moon, 2002; Moon, 2000; 탁지현, 2005)와 일관된 결과로서, 모든 연구에서 약 2/3 이상의 응답이 좋은 목소리와 좋은 사진을 연결하지 않았다.

실제로 인지실험 참여자들이 어떤 양상으로 좋은 목소리와 사진을 골랐는지를 자세히 살펴보는 것도 의미가 있는 일일 것이다. 다음의 <표 8>과 <표 9>는 각각의 목소리와 사진이 가장 좋은 목소리, 가장 좋은 사진으로 꼽힌 비율을%로 나타낸 것이다.

표 8. "가장 좋은 목소리" 선택도
Table 8. % of favorite voice chosen

화자성별	남자		여자	
세션종류	전신	얼굴	전신	얼굴
A	43	33	16	14
B	16	14	16	17
C	14	11	0	3
D	0	3	3	0
E	0	3	3	3
F	5	14	8	8
G	3	3	11	11
H	19	19	43	43
계	100%	100%	100%	100%

표 9. "가장 좋은 사진" 선택도
Table 9. % of favorite picture chosen

화자성별	남자		여자	
세션종류	전신	얼굴	전신	얼굴
P1	5	8	0	3
P2	3	0	22	9
P3	5	8	16	49
P4	16	8	5	6
P5	43	33	3	14
P6	16	19	0	3
P7	3	8	27	6
P8	8	17	27	11
계	100%	100%	100%	100%

세션별로 다른 사람들이 참여하였으나 소리의 경우는 세션에 상관없이 비슷한 양상으로 선호도가 나타나고 있다. 그러나 사진의 경우는 전혀 다른 모습을 보이고 있다.

<표 9>를 제시한 유일한 이유는 시각적인 정보에 의존하는 양상을 소리에 의존하는 양상과 비교하기 위한 것이다. 이로 인하여 혹시 생길지 모르는 개인의 권위침해를 방지하기 위하여 <표 9>에 나와있는 P1, P2, ... P8 은 본 논문의 앞에서 제시한 사진과 연관지을 수 없도록 전혀 다른 무작위 순으로 부여한 번호이다. 단 같은 번호의 사진은 전신사진세션이나 얼굴사진세션에서 모두 같은 사람을 가리키도록 재배열하였다. 예를 들면 남자 P1 은 전신이나 얼굴이나 같은 사람을 가리킨다.

<표 9>에서 볼 수 있듯이 소리에 대한 선호도와 달리 사진에 대한 선호도는 전신인지 얼굴인지에 따라 큰 차이를 보이고 있다. 이처럼 외모보다 목소리에 대한 선호도가 더 보편적일 수 있다는 결과는 선행연구(Moon, 2002; Moon, 2000; 탁지현, 2005)와 대체로 일치하였다.

본 연구의 결과를 같은 미국인의 자료를 한국인에게 들려주었던 탁지현(2005)의 결과와 비교해보자. 탁지현(2005)에서 한국인들은 목소리에 대하여 남자화자의 경우는 약 30%의 참여자가 본 연구와 같은 A 목소리를 선호하였고, 여자화자의 경우 70%가 본 연구와 같은 H 목소리를 선호하였다. 흥미로운 것은 여자화자의 경우는 한국인들의 H 선호도가 훨씬 높았던 반면, 남자화자의 경우는 A 에 대한 선호도가 훨씬 약하였다.

그러나 가장 좋은 사진의 경우 한국인의 선호도는 미국인과 전혀 달라서, 남녀화자 모두 전신사진세션이나 얼굴사진세션에서 전혀 다른 사람을 선호하는 것으로 나타났다. 한국인은 남자 전신사진으로는 P8(46%), 남자 얼굴사진으로는 P3(39%), 여자 전신사진으로는 P5(43%), 여자 얼굴사진으로는 P8(32%)을 가장 선호하여 미국인들과 극명한 대조를 이루었다. 이렇게 볼 때, 언어권에 따른 문화의 차이는 외모에 대한 판단에 큰 영향을 미치는 것처럼 보인다. 그러나 영어권이나 한국어권 모두 목소리의 선호도에서는 같은 목소리를 꼽았다는 것이 매우 흥미롭다. 이런 점에서 본다면 목소리가 외모보다 더 문화권의 차이를 뛰어넘는 보편성이 크다고 할 수 있을 것이다.

4. 결 론

본 연구는 한국어와 영어의 2 개 언어자료를 이용하여 같은 문화권과 다른 문화권 사람들로 하여금 목소리를 듣고 그 목소리의 주인공을 찾도록 하는 과정을 통하여, 다음의 세 가지 질문에 대한 답을 찾고자 하였다.

1. 그 목소리로 인하여 만들어지는 외모에 대한 인상이 어느 정도 정확하며, 어느 정도 보편적인가?

2. 신체의 시각적인 정보의 유무가 목소리를 듣고 시각적인 정보를 유추하는데 영향을 미치는가?

3. 목소리를 듣고 외모에 대한 정보를 만들어내는 데에 언어권을 포함한 문화적 요소가 영향을 미치는가?

영어자료를 미국인에게 들려준 본 연구의 결과를 영어자료를 한국인에게 들려준 탁지현(2005), 한국어자료를 한국인에게 들려준 Moon(2002), 같은 한국어자료를 미국인에게 들려준 Moon(2000)의 자료와 비교한 결과 다음과 같은 결론에 이르렀다.

1. 목소리로 인하여 만들어지는 외모에 대한 인상은 상당히 보편적으로서, 목소리를 들은 사람들은 비슷한 이미지를 떠올리게 된다. 그 떠올린 이미지가 실제의 이미지와 일치하지 않는 경우에도 다수의 사람들은 비슷한 이미지를 떠 올렸다.

2. 신체적인 특징이 드러나는 전신사진이 제시되었을 때에, 신체적인 특징이 확연하지 않은 얼굴확대사진이 제시되었을 때보다 더 정확하고 공통적인 반응이 야기되었다. 따라서 목소리는 그 주인공의 신체적인 요소의 일부를 반영한다고 생각할 수 있을 것이다.

3. 같은 목소리와 같은 사진을 본 결과가 이처럼 듣는 이의 국적에 따라 다르다면, 그것은 목소리를 듣고 그 주인공을 유추하는 과정에 듣는 이들이 속한 문화가 영향을 미친다는 것을 의미한다고 유추할 수 있을 것이다.

4. 목소리는 외모보다 더 보편적인 요소, 즉 더 많은 사람이 그 좋고 나쁨에 대해 공감할 수 있는 가능성이 있는 것 같다. 만일 이것이 사실이라면, 많은 사람들에게 호감을 주는 목소리를 합성하는 것이, 많은 사람들에게 호감을 주는 모습을 그려내는 것보다 훨씬 수월할 것이라고 짐작할 수 있을 것이다..

본 연구는 선행연구들과 함께 목소리와 그로 인하여 떠오르는 인상간에는 우연(여덟개의 선택 중 하나를 우연히 선택할 확률, 즉 1/8) 이상의 관계가 있음을 보여주었다. 그렇다면 추후에는 과연 목소리의 어떤 물리적인 특성이 어떤 인상과 관계가 있는 것인지를 규명하는 노력이 필요할 것이다. 그러한 노력이 결실을 맺는다면, 그것은 소리로서 필요한 인상을 만들어낼 수 있는 획기적인 발전을 이룩할 것이다.

감사의 글

본 연구에 녹음자료 및 사진자료를 기꺼이 제공해준 탁지현씨와, 본 연구의 인지실험실행에 장소 및 시설 제공 등 큰 도움을 주신 University of Massachusetts, Amherst 의 John Kingston 교수님께 깊은 감사를 드립니다.

참고문헌

Choi, Y. & Shin, M. (2004). "Evaluation Dimensions on the Impression of the Voice of Utterance," *Proceedings of 2004 Korean Society of Phonetics*, pp. 57-61.

(최양규 · 신명선. (2004). "발화 음성의 인상에 대한 평가차원," 2004 대한음성학회 가을 학술대회 발표논문집, pp. 57-61.)

Tak, J. (2005). "Voice and the image triggered by the voice–American speakers and Korean listeners," M.A. thesis, Ajou University.

(탁지현, (2005). "음성과 그로 인해 만들어지는 이미지의 연계성: 미국인화자와 한국인청자," 아주대학교 석사학위논문.)

Cohen, M. M., Beskow, J. & Massaro, D. W. (1998). "Recent developments in a facial animation: An inside view," *Proceedings of Auditory Visual Speech Perception '98*. pp. 201-206. Terrigal-Sydney Austrailia, December.

Fellowes, J. M., Remez, R. E. & Rubin, P. E. (1997). "Perceiving the sex and identity of a talker without natural vocal timbre," *Perception & Psychophysics,* 59, pp. 839-849.

Goggin, J. P., Thompson, C. P., Strube, G. & Simental, L. R. (1991), "The rold of language familiarity in voice identification," *Memory & Cognition,* 19, pp. 448-458.

Green, K. P., Kuhl, P. K., Meltzoff, A. N. & Stevens, E. B. (1991). "Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect," *Perception & Psyhophysics* 50 (6). pp. 524-536.

Leinonen, L. & Hiltunen, T. (1997). "Expression of emotional motivational connotations with a one-word utterance," *J. Acoust. Soc. Am.*, 102(3), pp.1853-1863.

Linville, S. E. (2000). "The aging voice," In Kent, R.D. & Ball, M.J (Eds.), *Voice Quality Measurement* (pp. 359-376). San Diego: Singular Publishing Group.

Kent, R. D. (1994). "The rainbow passage," (n.d.) Retrived July 1, 2004, from http://www.alt-usage-english.org/rainb_txt.html.

Kuhl, P. K. & Meltzoff, A. N. (1984). "The intermodal representation of speech in infants," *Infant Behavior and Development*, 7, pp. 361-381.

Kuhl, P. K. & Meltzoff, A. N. (1988). "The bimodal perception of speech in infancy," *Nature*, Dec. 10., pp. 1138-1141.

Moon, S-J. (2000). "Is what you hear what you see, even in a foreign language?" *J. Acoust. Soc. Am.* 108:5 Pt.2, p. 2482.

Moon, S-J. (2002). "What you hear is what you see?" *Journal of the Acoustical Society of Korea,* 21.1E, pp. 31-41.

Murray, I. R. & Arnott, J. L. (1993). "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *Journal of the Acoustical Society of America*, 93, pp. 1097-1108.

Remez, R. E., Fellowes, J. M. & Rubin, P. E. (1997). "Talker identification based on phonetic information," *Journal of Experimental Psychology,* 23, pp. 651-666.

Remez, R. E., Pardo, J. S., Piorkowski, R. L. & Rubin, P. E. (2001). "On the bistability of sinewave analogs of speech," *Psychological Science,* 12, pp. 24-29.

Scherer, K. R., Banse, R., Wallbott, H. G. & Goldbeck, T. (1991). "Vocal cues in emotion encoding and decoding," *Motivation and Emotion*, 15, pp. 123-148.

• **문승재 (Moon, Seung-Jae)**
아주대학교 인문대학 영어영문전공
경기도 수원시 영통구 원천동 산5번지
Tel: 031-219-2827 Fax: 031-219-1617
Email: moon@ajou.ac.kr
관심분야: 음성학, 심리언어학, 음악음향학

부 록 : The Rainbow Passage

When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

Throughout the centuries men have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky. Other men have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflections of the sun's rays by the rain. Since then physicists have found that it is not reflection, but refraction by the raindrops which causes the rainbow. Many complicated ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the water drops, and the width of the colored band increases as the size of the drops increases. The actual primary rainbow observed is said to be the effect of superpositioning of a number of bows. If the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green lights when mixed form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue (Kent, 1994).