음성과 그로 인해 만들어지는 이미지의 연계성

- 미국인 화자와 한국인 청자 -

탁지현, 문승재
아주대학교 영어영문학과

Voice and the Image triggered by the Voice

- American speakers and Korean listeners -

Ji-Hyun Tak, Seung-Jae Moon
Department of English Language and Literature, Ajou Univ.

yongsamo@hanmail.net, moon@ajou.ac.kr

Abstract

We can easily recognize the voices already known to us. But what about unknown voices? Is there any relationship between voices and the images triggered by the voices? Actually, this question has been partly addressed by Moon(2000, 2002). The current study aims at shedding some more lights on the topic by investigating the relationship between unknown foreign voices and the images triggered by them.

Speech samples from 16 American males and females (8 each) were recorded and 180 Korean subjects without any knowledge of the American speakers were asked to match the voices with the corresponding photos. And the number of corrects matches between voices and pictures of the current study was less than that of Korean-speakers and Korean-listeners case. But in terms of the majority matches, regardless of correctness, the present study showed a similar trend: that is, there is more than a chance relationship between voices and the images triggered by the voices.

I. 서론

본 연구의 목적은 영어 화자의 음성적 정보와 그 음성으로 인하여 화자를 전혀 모르는 한국인에게 떠오르는 화자의 이미지간에 관련성이 있는 지를 규명하는데 있다. 우리말 화자의 목소리와 그 목소리로 인하여 연상되는 외모와의 관계는 이미 Moon (2000, 2002)에서 한국인 화자-미국인 청자, 한국인 화자-한국인 청자간에 어느 정도 밝혀진바 있다. 본 연구에서는 외국인 화자에 대한 한국인 피험자들의 인지실험을 더함으로써, 지금까지의 결과에 보충적인 자료를 확보하고, 선행연구 결과와 비교하여 말소리와 그로 인해 생기는 이미지와의 관계를 규명하려고 한다.

본 연구에서 밝히고자 하는 것은, 지금까지 많이 연구되어온 "화자인식"과는 본질적으로 다른 성질의 것이다. 일반적인 화자인식은 이미 알고 있는 사람의 목소리를 듣고 화자를 인식하는 과정으로서, 이에 대해서는 이미 많은 연구가 진행되었다. 예를 들면, Remez(1997)는 화자인식은 음향적(acoustic)인 정보보다는 분절음적(segmental)인 정보에 의존하며, 말소리(natural utterance)가 아닌 단순한 SR(sinewave replication)음 만으로도 이미 알고 있는 화자를 인지할 수 있다고 주장한다. 그러나 본 연구에서 다루고자 하는 주제는 이처럼 이미 알고 있는 사람에 대한 화자인식이 아닌, 전혀 모르는 사람의 목소리와 그로부터 연상되는 화자에 대한 시각적 정보와의 관계이다

II. 실험

1. 화자

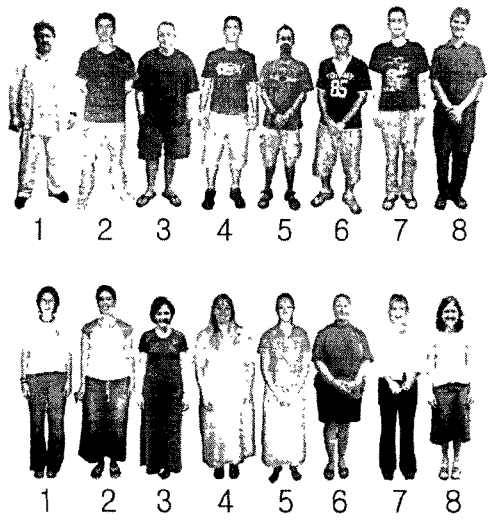
실험을 위한 자료는 우리말을 전혀 모르는 미국인 남여 8명씩 총 16명으로부터 수집하였다. 언어상의 차이를 고려하여 화자는 모두 백인으로 하였으며, 연령은 10세 이상 차이가 나지 않도록 하였다.

2. 자료

녹음을 위한 말 자료로는 언어치료분야에서 널리 쓰이는 "The rainbow passage"를 선정하였다. 말 자료는 화자들 스스로 자연스럽게 읽었다고 느낄 때까지 읽고난 후 녹음을 하였으며, 자연스럽게 읽으라는 것 외에 별다른 지시는 하지 않았다. 화자들의 녹음은 32kHz, 16kHz 등 여러 가지 sampling rate로 이루어졌다.[1]

실험에 필요한 이미지 자료를 위하여 녹음 직후 디지털 카메라(Sony Cybershot u-20)를 사용하여 사진을 찍었으며, 사진은 Adobe Photoshop(ver. 7.0.1)을 사용하여 전반적인 체격조건 등을 알 수 있는 전신사진과 그렇지 않은 얼굴사진의 두 세션<그림 1,2>으로 나누어 비율에 맞춰 편집하였다.

<그림 1> 전신세션



3. 인지실험

인지실험은 아주대학교 음성학 실험실에서 2대의 컴퓨터와 헤드폰(Inkel YH-3000과 GW-500M)을 사용하여 피험자 2명이 동시에 참여할 수 있도록 하였다. 인지실험에 참가한 피험자들은 미국인 화자를 전혀 모르고 청력에 이상이 없다고 스스로 판단한 아주대학교 학부생 180명(전신세션: 105명, 얼굴세션: 75명)으로

<그림 2> 얼굴세션

구성되었다. 음성파일은 피험자들이 직접 선택하며 들어보아야 했기 때문에 간단히 Praat(ver. 4.2.07)의 조작법을 설명한 후, 직접 음성파일(남자: MA~MH, 여자: GA~GH)을 선택하고 목소리에 맞는 사진을 고를 수 있도록 하였다. 음성파일을 듣는 시간이나 횟수 등에는 제한을 두지 않았다. 단, 중복하여 고르는 사진은 없도록 주지시켰다. 목소리와 사진을 연결하는 것 외에 가장 좋은 소리와 가장 좋은 인상도 하나씩 선택하도록 하였는데, 이것은 선행연구(Moon 2000, 2002)와 비교해 보기 위함이다.

III. 결과

1. 전신세션

전신사진에 관한 인지실험에 참가한 피험자들은 총 105명으로 다음 <표 1>은 남자 전신사진, <표 2>는 여자 전신사진에 해당하는 인지실험결과의 응답률을 백분율로 환산한 것이다.

[1] 이것은 마군 부대 등, 화자들의 환경 상 연구자가 직접 녹음할 수 없는 경우가 많아서 부득이 화자들에게 접촉이 가능한 사람에게 부탁하여 녹음이 이루어지는 과정에서 통제가 철저히 이루어지지 못하여 빚어진 상황이었다.

<표 1> 남자 전신사진 세션 결과

	목소리								
사진	MA	MB	MC	MD	ME	MF	MG	MH	계
1	16	18	15	6	1	8	21	15	100
2	17	9	9	8	5	22	6	26	100
3	19	25	16	15	4	10	7	5	100
4	5	9	10	12	29	11	15	9	100
5	12	12	6	24	26	7	9	5	100
6	2	10	10	18	20	10	26	4	100
7	5	3	27	11	10	14	12	18	100
8	24	14	8	6	7	18	5	19	100
계	100	100	100	100	100	100	100	100	

<표 2> 여자 전신사진 세션 결과

	목소리								
사진	GA	GB	GC	GD	GE	GF	GG	GH	계
1	3	7	19	22	10	11	9	20	100
2	7	38	14	7	7	9	10	9	100
3	10	7	8	6	8	25	28	10	100
4	8	18	9	28	16	4	13	5	100
5	7	8	12	14	10	10	11	30	100
6	3	11	7	17	43	5	10	5	100
7	5	10	24	2	6	18	19	16	100
8	59	2	8	4	1	19	2	6	100
계	100	100	100	100	100	100	100	100	

위의 표에서 색칠이 되어 있는 부분은 각 목소리에 대하여 피험자들이 가장 많이 선택한 사진의 빈도를 나타낸 것이며, 밑줄이 쳐있는 것은 목소리의 실제주인공을 나타낸 정답에 해당한다. 밑줄과 색칠이 되어 있는 MA-8, MB-3, MG-6, GA-8, GB-2, GF-3들은 목소리의 주인공임과 동시에 피험자들이 가장 많이 고른 사진이다. 목소리의 주인공(사진)을 제대로 선택한 정답은 남여 모두 8명 중 3명씩(남: MA, MB, MG, 여: GA, GB, GF)이다. 정답 여부와 관계없이 목소리에 따라 가장 많이 선택된 사진의 비율을 보면 남자는 평균 25%, 여자는 34%였다.

2. 얼굴세션

얼굴사진세션에 참가한 피험자들은 총 73명으로 그 결과는 <표 3, 4>에 나와 있다.

얼굴세션의 정답은 8명중 1명(남: MB-1, 여: GB-6)으로 전신세션에 비해 낮은 것으로 나타났다. 가장 많이 선택된 사진의 비율도 남자는 평균 25%, 여자는 26%로 전신세션보다 낮은 비율로 나타났다. 이처럼 얼굴세션의 경우가 전신세션의 경우보다 정답률이나 최다선택률에서 낮게 나타난 것은 Moon(2000, 2002)와 일치하는 현상으로서, 아마도 전신세션에 비해 화자의 체격이나 스타일등의 전반적인 정보가 얼굴로만 제한되어 있었기 때문인 것으로 풀이될 수 있을 것이다.

<표 3> 남자 얼굴사진 세션 결과

	목소리								
사진	MA	MB	MC	MD	ME	MF	MG	MH	계
1	21	24	5	9	1	8	9	21	100
2	21	15	5	12	17	12	11	7	100
3	11	7	9	17	11	15	24	7	100
4	5	13	3	17	43	7	9	3	100
5	8	12	9	13	11	20	16	11	100
6	5	17	24	13	12	12	3	13	100
7	21	11	17	5	1	13	16	15	100
8	7	1	27	12	4	13	12	24	100
계	100	100	100	100	100	100	100	100	

<표 4> 여자 얼굴사진 세션 결과

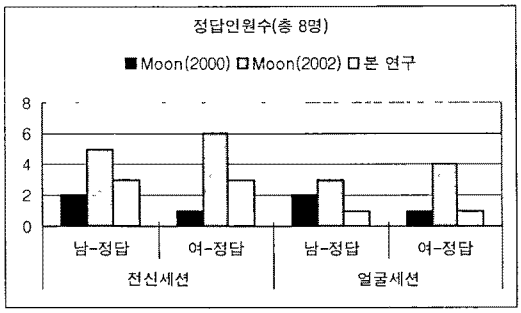
	목소리								
사진	GA	GB	GC	GD	GE	GF	GG	GH	계
1	5	12	10	22	14	12	11	14	100
2	14	5	22	4	10	21	15	10	100
3	10	11	12	10	4	11	22	21	100
4	21	7	4	19	3	14	12	21	100
5	22	18	16	7	10	7	10	11	100
6	5	25	21	4	4	25	10	7	100
7	1	18	10	12	47	5	4	3	100
8	22	4	5	22	10	5	16	15	100
계	100	100	100	100	100	100	100	100	

3. 선호하는 소리/ 선호하는 인상

본 연구에서는 가장 좋은 소리와 가장 좋은 인상을 연관지어 생각하는지 여부를 알기위한 실험도 함께 이루어졌다. 가장 좋은 소리와 가장 좋은 인상을 올바르게 연관지었던 비율이 30%, 그렇지 않은 비율이 70%나 나타난 것으로 보아 단순히 인상이나 목소리가 좋다고 하여 꼭 좋은 목소리나, 인상으로 연관지어 생각하지는 않았다. 이러한 결과는 Moon(2000, 2002)에서도 그렇지 않다고 판명된바 있다.

IV. 논의

<그림 3> 연구별 정답 수 비교



위의 <그림 3>은 화자-청자, 청자-화자간의 문화적 차이로 인한 세션별, 성별 정답 인원수를 비교하기 위하여 나타낸 것이다. 본 연구는 정답률에서는 남여 각 8명 중 5명이 정답이었던 한국인화자-한국인청자의 Moon(2002)와는 큰 대조를 보인반면, 목소리와 사진의 올바른 연결이 거의 없었던 한국인화자-미국인청자의 Moon(2000)과는 비슷한 양상을 보였다.

<그림 4> 연구별 정답률 및 최다선택률 비교

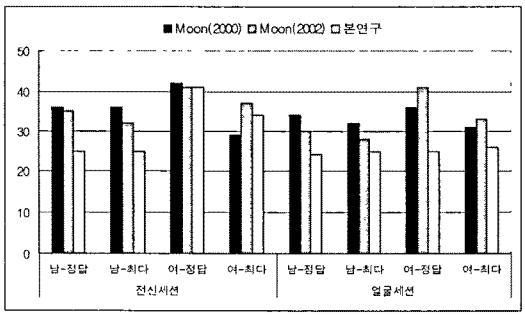


그림 4는 정답률(정답을 응답한 반응수의 백분율)과 최다선택률(가장 많은 사람이 고른 반응수의 백분율)을 나타낸 것이다. 정답수와는 달리 정답률과 최다선택률에 있어서는 다른 문화권을 대상으로 한 본 연구가, 역시 다른 문화권을 대상으로 한 Moon(2000)보다 같은 문화권을 대상으로 한 Moon(2002)와 더 비슷한 양상이라는 흥미로운 사실을 발견할 수 있다. 혹시 이 현상이, Moon(2000)에 참여한 미국인들은 한국어를 전혀 몰랐던 반면, 본 연구에 참여한 한국인들은 모두 대학생으로 영어에 노출이 많이 되었던 탓에 기인하는 것은 아닌지 생각해 볼 필요가 있다.

세 연구 모두 공통적으로 남자보다 여자 화자의 경우에 더 높은 정답률을 보였다. 아마도 이것은 여자들의 경우, 화장이나 다양한 옷차림 등, 외적인 특성이 남자들에 비해 풍부하여 피험자들이 선택하기가 훨씬 쉬웠던 것이 아닐까 추측해보게 된다. 특히 세 연구 모두 얼굴세션보다 전신세션의 정답률과, 최다선택률이 높은 것이 이처럼 외형적인 정보(체형, 스타일등)가 많은 역할을 한다는 추론을 뒷받침해준다.

Moon(2000)과 Moon(2002), 그리고 Moon(2002)와 본 연구를 각각 비교해 보면 같은 문화권의 화자-청자(Moon, 2002)일 때가 정답 및 최다선택률이 높았음을 알 수 있다. 이것은 문화에 따라서 듣는 목소리를 기대하는 모습이 다르다는 것을 의미한다고 할 수 있을 것이다. 이것은 사람의 얼굴에 대한 인상을 형성할 때 문화에 따라 서로 다른 차원을 사용한다는 주장[4]와 매우 밀접한 관계가 있다고 하겠다.

한 가지 특이한 것은, 본 연구에서는 피험자가 직접 목소리를 듣고 선택하는 화자의 목소리 배열과정에서 이미지와의 정답을 맞춘 경우가 두 세션 모두 초반(전신: MA, MB, MG, GA, GB, GF 얼굴: MB, GB)에 치중이 되어있었다는 것이다. 이것은 피험자들이 사진을 중복하여 고를 수 없는 점을 감안, 처음에는 정답을 고르기 위해 목소리를 신중히 여러 번 듣고 결정을 했기 때문에 후반 보다는 정답률이 높게 나타났던 것으로 해석할 수도 있을 것 같다. 그러나 이러한 순서에 따른 현상이 선행연구(Moon 2000, 2002)에서도 있었는지 확인한 결과, 선행연구에서는 그러한 현상이 발견되지 않았다. 따라서 이것을 단순히 순서에 의한 인위적인 결과라고 단정 지을 수는 없을 것이다 그러나 이후 연구에서는 순서의 영향도 고려하여 실험을 계획해야 할 것이다.

본 연구에서는 여자 화자 녹음 시 sampling rate를 균일하게 통제하지 못하였으나, 정답률이나 특히 최다선택률이 Moon(2000, 2002)에서와 비슷하게 나온 점으로 보아, 목소리를 듣고 화자를 선택할 때 sampling rate는 결정적인 영향은 끼치지 않았던 것으로 보인다. 그러나 이러한 기술적인 면도 향후 연구에서는 철저하게 통제되어야 할 부분이다.

V. 결론

세 연구의 비교를 통해 내릴 수 있는 결론은 다음과 같이 요약할 수 있다 첫째, 문화에 따라 정자의 화자에 대한 인지차이가 나타난다. 둘째, 성별로 보면 남자보다는 여자가, 세션별로는 얼굴보다는 전신세션에서 정답과 최다 선택사진이 많다. 셋째, 선호하는 목소리와 이미지와의 상호 연관성은 없다. 넷째, 위의 세 가지 결론은 비록 모르는 사람일지라도 화자에 대한 시각적 정보가 많을수록 높아질 수 있는 결과라는 것이다.

이것은 익숙하든 낯설든, 목소리와 그 목소리로 인하여 생성되는 인상 간에는 서로 관계가 있음을 의미하며, 나아가서는 원하는 인상을 어느 정도까지는 목소리로 만들 수 있음을 시사한다고 하겠다. 이러한 사실은 음성합성분야나 심리언어학분야 등 음성연관분야에 여러 가지로 기여를 할 수 있을 것이다.

참고문헌

[1] Moon, Seung-Jae, "Is what you hear what you see, even in a foreign language?" *The Journal of the Acoustical Society of America*, Vol. 108, No. 5, Pt. 2, November 2000

[2] Moon, Seung-Jae, "What You Hear is What You See?" *The Journal of the Acoustical Society of Korea*, Vol.21, No.1E, pp. 31~41, March 2002

[3] Remez, R. E., Fellowes, J. M., & Rubin, P. E. "Talker identification based on phonetic information", *Journal of Experimental Psychology: Human Perception and Performance*, 23, pp. 651-666, 1997

[4] 이경성, "한국 사람들은 사람들의 얼굴인상을 어떠한 차원들로 지각 하는가?", *Korean Journal of Social and Personality Psychology*, Vol.16 No.2, 2002