합성음 평가 방법 고찰1)

남기춘*, 구민모*, 김종진**, 최양규***
* 고려대학교 심리학과, **한국전자통신연구원, *** 춘해대학 유아특수치료교육과

Review on the synthetic sounds evaluation methods

Kichun Nam*, Minmo Goo*, Yanggyu Choi**
* Department of psychology, Korea University
** Electronics and Telecommunications Research Institute
*** Department of therapeutic & special education for children, Choonhae Colleage
kichun@korea.ac.kr

1. 합성음 평가의 필요성

컴퓨터를 이용해 인간의 행동을 모사하려는 노력이 여러 분야에서 줄기차게 이루어져 왔다. 잘 알려진 것처럼 1980년대에는 인공지능이라는 연구 분야가 활성화되어 인간의 인지 행동을 컴퓨터 프로그램으로 구현하려는 노력이 있었고, 1990대에는 인간의 뇌를 모사하여 인공시스템을 개발하려는 연구가 활발하게 이루어졌다. 음성과 관련된 연구에서도 인간의 음성 언어 행동을 모사하거나 인간 음성 언어 행동을 대신할 수 있는 컴퓨터 분야의 연구가 활발하게 이루어져 왔다. 본 연구에서는 인간의 음성 언어 행동 중에 음성 산출을 컴퓨터로 구현하려는 음성 합성(speech synthesis) 평가 연구를 다루려한다.

모든 분야의 연구 활동이 그렇듯이 음성 합성 연구에서도 연구의 핵심은 음성 합성 시스템의 설계와 구현과 현재 구현된 음성 합성 시스템이 지니고 있는 성능과 한계가 무엇인지를 조사하는 것이다. 즉, 시스템 자체를 개발하는 것과 시스템의 성능의 현실을 파악하고 부족한 면을 개선할 수 있는 단서를 제공하는 평가가 중요하다. 음성 합성 연구에서 지금까지는 전 세계적으로 음성 합성 시스템 자체를 개발하는데 주안점을 두고 합성 시스템의 성능과 한계를 평가하는 연구는 상대적으로 도외시 되어 왔다. 저자의 짐작으로는 음성 합성 기술 자체를 개발하는 것이 이전에 구현된적이 없었기 때문에 시스템 개발이 상당히 어려웠고 이에 따라 시스템 개발에 더 큰 비중을 두었던 것으로 생각된다. 음성 합성 시스템의 성능과 한계를 밝히는 연구가 점점 더 중요시 되는데, 그 이유는 현재 상태에서 여러 종류의 음성 합성 기술이 개발되었기 때문에 이제는 개발된 시스템이 어떤 문제점을 가지고 있으며 이런 문제점을 어떤 방법으로 개선하고 개선된 시스템을 어떻게 편리하게 사용할 수 있는지를 고려해야하기 때문이다.

음성 합성 시스템의 평가 연구는 주로 다음과 같은 이유 때문에 중요하다고 생각된다. 첫째, 음성 합성 시스템을 특정 상황에 응용하려고 할 때 현재 구현된 음성 합성 시스템이 사용하려는 상황에 적절한지를 평가해야 하기 때문이다. 예를 들어, 정보 자체만을 음성으로 전하는 것이 주목적일 때에는 정보를 정확히 전할 수 있는 능력이 요구되고 반면에 어린아이에게 음성으로 동화를 읽어 주는 시스템에 적용하려고 하는 경우에는 동화를 듣는 아이가 싫증나지 않도록 음성 합성 시스템을 구현하는 것이 중요하다. 또한, 정상적인 청력을 가지고 있는 사용자를 대상으로 할 때와 청각 시스템에 일부 손상을 입어서 정상적인 음성 정보처리가 어려운 사용자를 대상으로 하는 경우에도 분명히 다른 기능을 가진 음성 합성 시스템이 적용되어야 한다.

두 번째 이유는 현재 구현된 시스템의 성능평가를 위해 필요하다. 개발된 시스템이 얼마나 안정적인지를 평가해야 하고 어떤 음성 합성 시스템이 어떤 것보다 우수한지를 판단해야 하기 때문이다.

세 번째 이유는 좀 더 개선된 음성 합성 시스템의 개발

1) 본 연구는 정부통산부 위탁연구과제(1010-2002-0012)의 지원으로 수행 되었음

을 위해서는 현재 개발된 시스템이 지니고 있는 한계를 밝혀야 하기 때문이다. 시스템 개발 측면에서도 현재 지니고 있는 한계점을 지적할 수 있고 사용자 측면에서 한계점을 지적할 수도 있다. 그렇지만 결국은 음성 합성 시스템을 인간이 이용하는 것이기 때문에 사용자가 불편하게 느끼는 문제점을 찾아내고 이런 문제점을 해결하는 것이 더 중요하다고 생각된다.

네 번째로 음성 합성 시스템 평가가 중요한 이유는 인간이 음성을 산출하는 방법과 유사한 방법을 따르는 음성 합성 시스템을 개발하기 위해서이다. 인간이 만들어 내는 음성만큼 인간이 정보처리하기에 효율적이고 완벽한 것은 없다. 현재 개발된 음성 합성 시스템은 canned speech, 규칙 기반, 코퍼스 등을 기반으로 음성을 합성한다. 현재까지 개발된 시스템은 대부분이 코퍼스를 이용하는 방법인데 인간이 이런 방법을 반드시 따를 지는 의문이다. 입력된 문자를 해독하여 음운 정보로 바꾸는 일은 극히 제한된 상황에서의 음성 합성이다. 또한, 저장되어 있는 코퍼스의 특정 정보를 인출하여 합성하고자 하는 음운 정보를 산출하는 것도 극히 제한된 방법이다. 좀 더 일반적이고 모든 상황에 적절한 음성 합성 시스템을 개발하기 위해서는 인간이 어떤 방법으로 음성을 산출하는지를 이해해야 하고 인간의 음성 산출 원리를 음성 합성 시스템 개발에 적용해야 한다. 이렇게 하기 위해서는 음성 합성 시스템의 음성 합성 원리와 인간의 음성 합성 원리를 비교하여 음성 합성 시스템이 부족한 측면을 개선해야 한다.

이처럼 합성음 평가가 매우 중요한데 문제는 어떤 방법을 사용해야 음성 합성 시스템의 적절성, 성능, 한계점을 정확하고도 객관적으로 평가할 수 있는가이다. 다음 절에서는 기존에 개발되어 있는 합성음 평가 방법을 고찰하고 새롭게 시도되고 있는 인지신경과학적 합성 음성 평가 방법을 소개하려한다.

2. 기존에 개발된 합성음 평가 방법

기존에 개발된 합성음 평가 방법은 크게 세 종류로 대분된다. 이 세 종류를 소개하면 다음과 같다. 첫 번째 방법은 선호도 평가 방법이다. 선호도 평가 방법은 두 종류의 합성음을 들은 다음에 어느 것이 듣기에 더 좋은 지를 평가하는 것이다. 합성음이 쌍으로 제시될 때 어느 것이 더 명료한지 혹은 자연스러운지를 평가하는 경우도 있고 하나의 합성음을 들려준 후에 7점 scale 혹은 5점 scale 명료도나 자연성을 평가하게 하는 방법이다. 이 방법에서는 흔히 간단한 선호도 평정치를 이용하기도 하고 정신물리학적 scaling (psychophysical scaling)을 사용하기도 한다 (Goldstein, 1995; lawson, 1982; Pavlovic, 1990)). 선호도 평가 방법은 시행하기는 쉬우나 평가자의 주관, 비교가 되는 합성음의 쌍, 기준의 자극의 특성 등에 따라서 변할 수 있기 때문에 객관적인 지표라고 보기 어렵다(.

두 번째 방법은 합성음 청취 실험을 시행하는 경우이다. 연구 방법은 합성된 여러 언어학적 단위(예를 들어, consonants, vowels, rhymes, vowel-consonant transitions 등)를 제시하고 방금 들은 것이 무엇인지를 직접 보고하게 하거나 여러 종류의 후보 중에 하나를 표시하는 방법으로 이루어진다(Benoit, Grice, & Hazan, 1996; Goldstein, 1995; Logan et al., 1989). 대개 이런 방법은 명료도 평가에 자주 사용된다. 이런 명료도 평가를 통해 현재 개발된 음성 합성 시스템이 어떤 종류의 음을 잘 산출하지 못하는지를 평가하여 개발되어 있는 음성 합성 시스템의 수정 보완 및 새로운 시스템의 개발에 이용되고 있다. 사용되는 언어학적 단위는 어휘 이하의 단위에서 주로 사용된다. 미국을 비롯한 선진국에서는 여러 종류의 합성음 청취 자극 세트를 개발하여 사용하고 있다. 국내에서는 이런 평가 세트도 아직 개발되어 있지 않은 상황이라서 합성음 평가에 어려움이 많다. 이와 같은 종류의 청취 실험 방법은 여러 종류의 장점을 가진다. 우선, 상대적으로 실험이 간단하고 따라서 평가를 수행하기가 수월하며, 합성음 평가자의 반응이 분명하여 자료 분석이 용이하다. 또한, 합성음이 지니는 여러 종류의 분절(segment)의 특성을 쉽게 찾아 낼 수 있다는 것이 장점이다. 그러나 이 방법 또한 다른 방법처럼 여러 종류의 단점도 지니고 있다. 예를 들면, 이 방법은 어휘 이하의 분절과 같은 언어학적 단위에 대한 평가는 가능하나 prosody와 같은 suprasegmental unit, 문장, 덩이글(text) 등과 같이 단위가 크고 복잡한 언어학적 단위를 평가하기에는 어려움이 있다는 것이다. 또한, 이 방법의 약점은 명료도를 평가하는 데는 유리하지만 자연성이나 좀 더 상위 수준의 언어정보처리 특성을 조사하기에는 적절하지 않다는 것과 청취 결과를 모든 정보처리 후에 보고하기 때문에 평가자의 반응 준거 등에 의해 평가 결과가 달라질 수 있다는 것이다.

근래에 들어 인지심리학의 이론과 연구 방법을 응용하여 합성음을 평가하려는 시도가 보고 되고 있다 (Delogu, Conte, & Sementina, 1998). 예를 들어 Pisoni 등(Pisoni & Hunnicutt, 1980)은 자연음 문장과 합성음 문장을 평가자에게 제시한 후에 제시된 문장을 이해하는 정도를 측정하였다. 실험 결과는 합성음 조건이 자연음 조건에 비해 개개의 단어와 지엽적인 정보를 기억하고 이해하는 정도에서는 더 우수하였지만 문장 전체의 의미를 파악하게 하였을 때에는 자연음 조

건에서 더 우수 하였다. 또한, Delogu 등 (1998)은 합성음 평가에 영향을 줄 수 있는 인지적 요인(예를 들어, 주의, 지각, mental load 등), 과제 특성(single task or dual task) 등을 이용하여 합성음 문장과 자연음 문장을 비교하였다. 이런 방법은 객관적이고 좀 더 심도 깊게 합성음으로 구성된 자극에 대한 언어정보처리 특성을 밝힐 수 있어서 유리하다고 생각한다. 문제는 이런 방법을 이용한 평가가 근래에 이루어지기 시작했기 때문에 인지심리학이나 언어심리학 등에서 밝혀진 현상이나 이론을 중심으로한 연구 패러다임이 다양하게 개발되어 있지 않다는 것이다.

이제까지 기존에 개발된 합성음 평가 방법을 개략적으로 살펴보았다. 위에서 살펴보았듯이 기존의 방법들은 장점도 많지만 단점도 많다. 특히 기존에 개발된 합성음 평가 방법의 문제점은 때로는 평가자의 주관에 따라 평가 결과가 수시로 바뀔 수 있으며, 주로 분절 단위의 명료도를 평가하는데 적절하고 문장이나 더 큰 언어학적 단위와 합성음의 자연성 등을 다루기에는 부적합하다는 것이다. 또한, 기존의 평가 방법은 주로 off-line 과제로 합성음을 듣는 동안 무슨 일이 일어나는지는 알려 주지 못하고 단지 모든 정보처리가 끝난 다음에 나타나는 결과만을 알 수밖에 없다는 것이다. 이처럼 정보처리 결과를 보고하게 하는 경우에는 평가자의 추측 혹은 주관적 판단에 따른 잘 못된 결과를 얻을 가능성도 있다. 다음 절에서는 이런 여러 기존의 평가 방법이 가지는 객관성의 결여, off-line 과제의 결점, 얕은 수준의 정보처리 결과만 보여 주고 좀 더 상위 수준의 정보과정을 보여 주지 못하는 결점, 명료도와 자연성을 함께 평가할 수 없다는 결점 등을 극복할 수 있는 여러 종류의 인지신경과학적 연구 방법을 소개하려한다.

3. 인지신경과학적 합성음 평가 방법

인지신경과학이라는 분야가 19990년대 후반에 들어 전 세계적으로 급속히 부상하고 있다. 이 분야에서는 인지심리학, 신경과학, 의학, 전산학, 언어학 등의 다양한 분야가 함께 녹아들어 인간의 인지 행동이 뇌의 구조 혹은 신경들의 connection network과 어떤 관련을 맺고 있는 지를 밝히는 분야이다. 인지기능을 뇌와 관련시켜 연구할 수 있게 된 계기로 새로운 뇌의 활동을 측정할 수 있는 새로운 기술의 개발과 인지 기능을 설명할 수 있는 이론적 토대를 생각할 수 있다. 특히 근래에 들어 뇌의 활동을 측정할 수 있는 EEG/ERP, fMRI, PET, MEG, SPECT 등이 개발되면서 뇌와 인지에 관한 연구가 활발하게 진행되었다. 전통적으로는 뇌를 손상(lesion)시키거나 손상된 뇌를 지닌 환자를 대상으로 이루어져 왔지만 그 진전이 매우 느려서 뇌의 복잡한 기능을 밝히기에는 부족한 면이 있었다. 그러나 근래에 개발된 기술로 인해 뇌를 손상시키지 않고도 뇌의 다양한 활동을 측정할 수 있게 되었고 이런 기술 개발이 인지신경과학이라는 새로운 학문분야를 탄생시켰다 (Gazzaniga, Ivry, Mangun, 1998). 이번 논문에서는 현재 고려대학교 인지신경과학연구실에서 수행하고 있는 합성음의 자연성과 명료도 평가 연구를 기초로 하여 인지신경과학적 방법을 소개 하겠다.

합성음의 자연성과 명료도를 측정하기 위해 몇 종류의 인지신경심리학적 방법을 사용할 수 있겠다. 먼저 합성음의 자연성을 평가하기 위해 심리측정론적 방법(psychometrical method)과 인지신경심리학적 방법(cognitive neuropsychological method)을 사용한다. 심리측정론적 방법은 평가자로 하여금 자연음과 합성음을 듣고 60여개로 구성된 항목에 대해 평가자가 주어진 자극에 대해 느끼는 심리적 상태를 평가하게 하고 평가된 scale상에서의 점수를 이용해 중요한 요인을 분석해낸다. 흔히 원점수(raw score)에서 중요한 차원인 요인(factor or dimension)을 찾아내기 위해 통계적인 기법인 요인분석(factor analysis)을 실시한다. 그 다음에는 자연음을 들었을 때와 합성음을 들었을 때 어떤 차원 혹은 요인에서 차이가 있는 지를 평가한다. 이런 의미미분법(semantic differentiation)과 요인분석을 통해 자연성을 구성하고 있는 심리적 요인을 추출할 수 있고 이렇게 추출된 요인들을 이용해 합성음의 자연성을 평가할 수 있다. 자연성을 평가하는 또 다른 방법은 EEG, ERP, fMRI를 이용하는 방법이다. 의미미분법과 요인분석을 통해 구한 도구를 사용해 자연음과 합성음이 어떤 심리 차원에서 차이가 있는지를 밝히고 이런 차이가 뇌파와 대뇌 활성화에 어떤 차이를 가져오는 지를 평가하는 방법이다. EEG와 ERP는 대뇌에서 신경활동 때문에 일어나는 전압 변화를 측정하는 것이다. 인지신경심리학분야에서는 EEG와 ERP를 이용한 연구가 1980년 후반부터 활발히 이루어져서 현재는 심리상태와 인지과정에 관련된 많은 기초 연구가 이루어져 있다 (Brown, & Hagoort, 1999). EEG와 ERP를 이용해 어떻게 이용할 수 있는가를 여기에서 논하기에는 지면이 너무 부족한 것 같다. 관심 있는 연구자는 위의 참고문헌을 참조하면 좋겠다. 결론적으로, 현재 진행되고 있는 자연성 평가 연구는 자연성을 구성하는 하위 요인들을 찾는 연구와 하위 차원에서 나타나는 심리상태와 관련된 신경생물학적 자표를 찾는 연구로 이루어져 있다.

합성음의 명료도를 평가하기 위해서도 여러 종류의 방법을 사용하고 있다. 한 가지 방법은 지능 검사나 성격 검사처럼 명료도를 표준적으로 나타낼 수 있는

norm을 개발하는 것이다. 즉, 명료도를 나타내는 통계 분포 상에서 현재 들은 합성음이 어느 위치에 있는가를 평정하고 이 평정 값을 표준 점수화하여 나타내는 방법이다. 그러나 이 방법은 비용과 노력이 너무나 많이 요구되어서 현재는 진행하지 않고 좀 더 합성음 평가에 대한 자료를 수집한 다음에 실시할 예정이다. 또 다른 명료도 평가 방법은 기존에 언어심리학과 인지심리학에서 개발한 패러다임을 이용하는 방법이다. 흔히 이 방법에서는 청취 자극을 평가자에게 들려 주고 어휘판단(lexical decision)과 같은 인지과제를 시킨다. 이런 인지과제는 주로 평가자가 합성음을 듣고 평가하는 동안 일어나는 정보처리과정을 측정하게 된다. 흔히 측정하는 종속 변인은 반응시간, 실수율, 기억정도, 주의가 필요한 정도, 추론과정 등을 평가한다. 이전의 인지심리학적 연구를 살펴보면, 합성음을 듣고 그 소리가 무엇인지를 판단하는 데에는 자연음과 큰 차이가 없지만 음성을 듣고 좀 더 상위 수준의 정보처리에는 문제가 있는 것으로 알려져 있다. 따라서, 기존에 인지심리학과 언어심리학에서 알려진 현상을 중심으로 합성음을 정보처리할 때 불리하게 작용하는 상위정보처리 과정을 밝힐 수 있다. 또 다른 명료도 평가 방법으로 ERP와 fMRI를 이용하여 자연음을 들었을 때와 합성음을 들었을 때 대뇌에서 어떤 변화를 수반하는지를 조사하는 연구를 진행하고 있다. 이미 언어정보처리와 관련하여 나타나는 뇌파의 양상과 뇌 활성화 영역에 관한 연구가 많이 진전되어 있다. 이런 자료를 기초해서 합성을 들을 때 나타나는 대뇌의 현상을 해석해 낼 수 있다. 명료도 평가 방법에 관한 내용을 정리해보면, 첫 번째는 인지심리학과 언어심리학에서 개발된 인구 패러다임과 현상을 이용하여 합성음의 명료도와 상위 수준의 정보처리 양상을 조사하고 이런 연구를 기반으로 하여 대뇌에서 합성음을 들을 때 변하는 인지요인을 찾는 연구를 진행하고 있다.

참고문헌

1. Benoit, C., Grice, M., & Hazan, V., (1996). The SUS test: A method fot the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, 18.
2. Brown, C. M., & Hagoort, P., (1999). *The neurocognition of language*. Oxford University Press.
3. Delogu, C., Conte, S., & Sementina, C., (1998). Cognitive factors in the evaluation of synthetic speech. *Speech communication*, 24.
4. Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R., (1998). *Cognitive neuroscience: The biology of the mind*. W. W. Norton & company: New York.
5. Goldstein, M., (1995). Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listeners. *Speech communication*, 16.
6. Lawson, G. D., (1982). Magnitude estimation of degraded speech quality by normal- and impaired-hearing listeners. *Journal of the Acoustical Society of America*, 72, 6.
7. Logan J. Green B., & Pisoni D., (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, 86, 2.,
8. Pavlovic, C. V., (1990). Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis systems. *Journal of the Acoustical Society of America*, 87, 1.
9. Pisoni, D., & Hunnicutt, S., (1980). Perceptual evaluation of MITalk: The MIT unresticted text-to-speech system. *Proceedings of ICASSP*, 80, 3.