단어빈도와 단어규칙성 효과에 기초한 합성음 평가

남기춘*, 최원일*, 이동훈*, 구민모*, 김종진**

* 고려대학교 심리학과

** 한국전자통신연구원

The text-to-speech system assessment based on word frequency and word regularity effects

Kichun Nam*, Wonil Choi*, Donghoon Lee*, Minmo Koo*, Jongjin Kim**

* Department of Psychology, Korea University, ** ETRI

Abstract

In the present study, the intelligibility of the synthesized speech sounds was evaluated by using the psycholinguistic and fMRI techniques. In order to see the difference in recognizing words between the natural and synthesized speech sounds, word regularity and word frequency were varied. The results of Experiment1 and Experiment2 showed that the intelligibility difference of the synthesized speech comes from word regularity. There were smaller activation of the auditory areas in brain and slower recognition time for the regular words.

I. 서론

합성음을 이용한 응용 시스템이 다양한 분야에서 사용되고 있다. 예를 들어, 전화기 자동응답시스템, 공공장소에서의 음성 안내시스템, TTS(text-to-speech system), 자동차 운행 음성 안내 시스템, e-mail 낭독 시스템 등의 분야에서 사용된다. 음성 합성 시스템을 특정 응용 시스템에 적절하게 사용하기 위해서는 응용도메인의 특성 분석과 이런 응용 도메인에 적합하도록 음성 합성 시스템을 개발하는 것이 필수적이다. 응용 도메인에 따라 음성합성 시스템이 갖추어야 할 요건이 달라지지만 최종적으로 그 시스템이 적절한지는 사용자의 판단에 의해 결정된다. 이런 측면에서 합성음을 사용자 관점에서 평가하는 것은 매우 중요한 과제이다. 현재 개발되어 있는 음성 합성기는 여러 측면에서 부족하다. 예를 들어, 현재의 시스템은 소음이 있거나 반향 상황에서도 이해하기 쉬운 합성음을 생성해 내지 못하며, 감정이나 전하려고 하는 의미에 따른 운율 조절의 어려움을 가지며, 또한 특정 화자의 말하는 스타일과 음성의 품질을 자유스럽게 조절하는 데에서도 어려움을 가진다. 본 연구는 현재 개발되어 있는 음성 합성기가 이처럼 여러 한계점을 지니고 있지만 구체적으로 현재의 합성기 기술로 생성된 합성음이 인간언어정보처리에 어떤 영향을 주는 지를 조사하기 위해 실시되었다.

기존의 합성음 평가 방법은 주로 off-line 과제를 이용한 청취실험이었다(Goldstein, 1995). off-line 청취실험에서는 정보처리과정 중의 합성음 이해 과정을 다루기보다는 모든 정보처리가 종료된 후에 나타난 결과를 평가한다(Delogu, Conte, & Sementina, 1998; Goldstein, 1995; Pisoni & Hunnicutt, 1980). 이런 off-line 청취실험과제는 일반적으로 사용하기 간편하고 기존에 개발된 여러 척도를 이용할 수 있다는 측면에서 장점을 지닌다. 그러나 이런 off-line 합성음 평가 방법은 합성음을 듣는 동인에 일어나는 인간언어정보처리 특성을 조사할 수 없고 또한 때로는 합성음 평가자의 주관에 따라 평가 결과가 달라질 수 있다는 단점을 지니고 있다. 본 연구에서는 이와 같은 기존의 off-line 합성음 평가 방법의 한계성을 극복하기 위해 언어심리학(psycholinguistics)에서 흔히 사용하는 어휘 판단 과제

(lexical decision task)를 사용하여 합성음으로 생성된 단어를 정보처리동안 나타나는 특성을 밝히고자 하였다.

인간의 언어정보처리 특성을 연구하는 언어심리학 연구에서는 음성, 음운, 어휘, 문장, 글 등의 언어학적 단위를 대상으로 한다. 본 연구에서는 이런 여러 종류의 언어정보 단위 중에서 단어 재인과정에 초점을 두었다. 단순히 자연음 어휘와 합성음 어휘를 듣고 어휘 판단하는 속도나 실수율을 조사하는 것만으로는 합성음 어휘 정보처리 특성을 충분히 조사하기에 부족해서 단어빈도(word frequency)와 단어규칙성(word regularity) 효과를 중심으로 연구를 수행하였다. 단어빈도 효과는 단어를 인식할 때 자주 사용되는 단어일수록 더 빨리 인식되는 현상을 의미한다. 기존의 언어심리학 연구에 따르면 단어빈도효과는 대뇌(brain)에 저장되어 있는 심성어휘집(mental lexicon)에서 해당되는 어휘를 접근(lexical access)하는데 영향을 주기 때문으로 알려져 있다(Balota, 1994 참조). 또한, 단어 규칙성 효과는 문자와 음운간의 불규칙적인 어휘를 이해하는 시간이 규칙적인 단어를 이해할 때 소요되는 시간보다 긴 현상을 의미한다. 단어규칙성효과도 단어빈도효과처럼 어휘를 심성어휘집에서 찾는데 관여하는 변인으로 알려져 있다(이광오, 1996). 국내에서 개발된 음성 합성시스템이 대부분 규칙 기반(rule-based)이기보다는 말뭉치(corpus)를 이용하고 있어서 단어빈도와 단어규칙성이 민감하게 작용할 것으로 생각된다.

본 연구는 두 종류의 실험 연구로 구성되어 있다. 실험 1은 언어심리학적인 연구 방법을 사용하여 음의 종류, 단어빈도, 단어 규칙성이 음성 단어 재인에 어떤 영향을 미치는지를 어휘판단과제를 이용해 조사하였다. 실험 2는 실험 1에서 사용된 변인에 따라 대뇌 활성화 영역이 어떻게 차이 나는지를 조사하였다.

II. 실험 1

실험 1은 두 가지 목적을 가지고 있다. 첫째는 단어를 재인할 때, 합성음의 음질의 수준에 따라 청취어휘정보처리에서 차이가 있는가를 알아보는 것이다. 둘째는 합성기에서 만들어진 소리와 인간의 음성을 들을 때 나타나는 현상이 동일하게 나타나는지를 알아보고자 하였다. 실험 1에서 사용된 과제는 기존의 합성기 평가 연구에서 사용되지 않은 청각 어휘 판단 과제(auditory lexical decision task)이다. 만약 음의 종류에 따라 음질이 달라지고 이것이 단어 재인에 영향을 미친다면 반응시간에서 차이가 날 것이라고 예측할 수 있다. 또한 시각적 단어재인이나 청각적 단어재인 연구에서 나타나던 단어빈도효과(Luce, 1986; Savin, 1963)나 단어규칙성 효과(Bauer & Stanovich, 1980; Parkin, 1982)가 합성음에서도 동일하게 나타난다면 각 조건간의 반응 시간의 양상이 유사할 것이라 예측할 수 있을 것이다.

2.1. 방법

2.1.1. 피험자 서울 시내 대학에 재학 중이고, 청력에 문제가 없는 대학생 21명이 실험에 참여하였다.

2.1.2. 실험 재료 및 설계 실험에 이용된 자극은 빈도에 따라 고빈도와 저빈도, 단어 규칙성에 따라 규칙과 불규칙으로 나뉘어져서 각 조건당 30개씩, 모두 120개의 어휘가 사용되었다. 그리고 이 어휘들은 음 종류 조건에 따라 크게 3가지 종류로 만들어졌다. 첫째, DAT 녹음기로 중년의 여성이 녹음한 어휘, 둘째, 고품질의 합성기를 이용하여 생성된 어휘, 그리고 마지막으로 저품질의 합성기를 이용하여 생성된 어휘였다. 어휘들은 모두 2음절의 한국어였고, "물가"(고빈도 불규칙단어), "방침"(고빈도 규칙단어), "물증"(저빈도 불규칙단어), "변심"(저빈도 규칙단어)과 같은 자극들이 사용되었다.

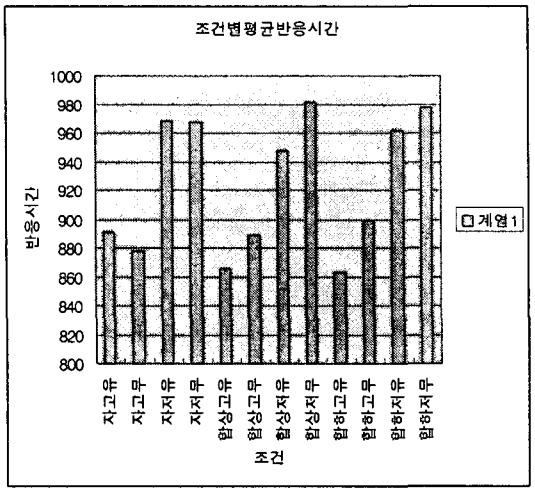
이 실험은 삼요인 피험자내 설계를 사용하였다. 첫 번째 독립변인은 3개의 수준을 갖는 음 종류 조건이고, 두 번째 독립 변인은 2개의 수준을 갖는 빈도이며, 세 번째 독립 변인은 2개의 수준을 갖는 규칙성이다. 종속변인은 각 조건별 평균 반응 시간으로 ms 단위로 측정되었다.

2.1.3. 절차 이 실험에 사용된 과제는 어휘 판단 과제이다. 어휘 판단 과제란 제시되는 자극이 단어인지 아닌지 최대한 빠르게 판단하는 것이다. 모니터에 "***"과 같은 부호가 800ms동안 제시된 후 500ms의 시간 간격이 있은 후에 피험자의 귀에 자극이 제시된다. 그러면 피험자는 그 자극을 듣고 단어인지 아닌지 판단해서 적절한 반응키를 눌러야 한다. 이러한 일련의 과정이 하나의 시행이고, 이러한 시행들이 반복되었다.

2.2. 결과 및 논의

각 실험조건의 대푯값(central tendency)을 추출하기 위해 피험자의 반응시간의 중앙값(median)을 통계 치로 사용하였고 각 조건의 평균 반응시간을 표 1에 나타내었다.

표 1에 나타난 조건별 반응시간을 가지고 삼원 피험자내 변량분석을 실시하였다.



<그림 1> 조건별 반응시간

그 결과, 먼저 음 종류에 대한 주효과(main effect)를 살펴보면 음 종류에 따른 주효과가 나타나지 않은 것을 볼 수 있다[F(2,20)=.311, p=.734]. 이는 적어도 어휘를 판단하는 표면 단계에서는 자연음과 합성음에서 차이가 나지 않는다는 것을 알 수 있다. 빈도에 대한 주효과와 규칙성에 따른 주효과는 모두 통계적으로 유의미한 결과를 나타내었다[F(1,20)=216.445, p<.05; F(1,20)=8.721, p<.05]. 즉 빈도가 높은 단어일수록 반응시간이 빨랐고, 음변화가 있는 규칙 단어일수록 반응시간이 빠르게 나타났다. 그리고 음 종류와 규칙성간의 상호작용효과 역시 통계적으로 유의미하게 나타났다[F(2,40)=6.121, p<.05]. 이 상호작용효과를 해석하기 위해서 각 음 종류에 따른 규칙성효과를 t-test를 통해 검증하였다. 그 결과 자연음 조건에서는 규칙성에 따른 반응 시간의 차이가 유의미하지 않았지만[t(41)=.680, p>.05], 고품질의 합성기음과 저품질의 합성기음의 경우에는 규칙성에 따른 반응시간의 차이가 통계적으로 유의미하였다[t(41)=-3.257, p<.05, t(41)=-3.172, p<.05]. 결국 자연음 조건에서는 규칙성 효과가 나타나지 않은 반면에 합성기 조건에서는 규칙성 효과가 나타났고, 음변화가 있는 조건에서 더 빠른 반응 시간을 나타냈다. 그리고 합성기의 품질에 따른 차이는 보이지 않았다.

III. 실험 2

실험 2는 합성음을 이해하려고 노력할 때와 자연음을 이해하려고 노력할 때 활성화되는 대뇌 영역을 조사하여 자연음과 합성음의 차이를 밝히려고 실시되었다. 대뇌 영역의 활성화를 조사하기 위해 fMRI 기법을 사용하여 하였다.

3.1. 방법

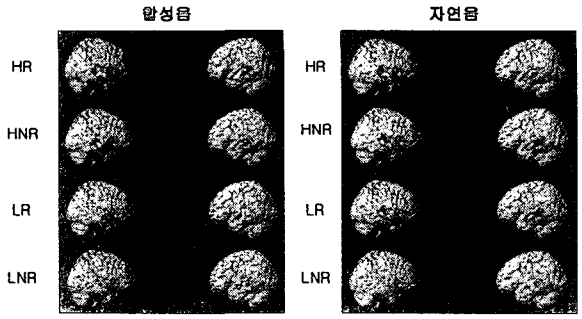
3.1.1. 피험자 피험자로는 신경학상 이상이 없고 오른손잡이인 남자 2명과 여자 2명이었다.

3.1.2. 실험 재료 및 설계 실험 1에서 사용하였던 실험 재료를 사용하였다. 실험에서 조작한 변인은 음의 종류(자연음 혹은 합성음), 단어 빈도(저빈도 혹은 고빈도), 단어규칙성(규칙단어 혹은 불규칙 단어)으로 2x2x2 설계였다.

3.1.2. 실험 절차 MR시스템은 한국과학기술원 fMRItpsxj에 있는 ISOL 3.0T forte를 사용하였다. 30초간의 활성기와 휴식기로 구성된 block 디자인을 사용하였다. 자연음과 합성음은 활성기에 제시되었다. 피험자는 활성기에 제시되는 단어를 듣고 그 단어의 의미가 무엇인지를 이해하라고 지시하였으며 다른 특별한 과제를 수행하지는 않았다.

3.2. 결과 및 논의

자연음과 합성음을 이해하는 동안 활성화되는 대뇌 영역 사진이 그림 2에 제시되어 있다.



HR: High frequency Regular word (ex: 목표, 공부, 분석)
HNR: High frequency Non Regular word (ex: 특립/등닙/, 단안/부란/, 현학/처학)
LR: Low frequency Regular word (ex: 공립, 엄단, 검직)
LNR: Low frequency Non Regular word (ex: 단중/단중/, 직함/지캄, 격현/경년)

그림 2에서 볼 수 있듯이 고빈도단어일 경우에는 규칙단어이든 불규칙단어이든 관계없이 자연음과 합성음간에 차이가 없었다. 자연음이든 합성음이든 1차 및 2차 청각피질에서의 활성화가 있었으며 또한 중측두피질에서도 유의미한 활성화가 있었다. 그러나, 저빈도단어의 경우에는 다른 결과가 나타났다. 자연음의 경우에는 음변화가 없는 단어의 경우, 고빈도 단어의 활성화와 거의 동일한 활성화가 나타났으나 음변화가 있는 경우에는 활성화수준이 떨어졌다. 반면에 합성음의 경우에는 고빈도 단어에 비해 전반적인 활성화수준이 떨어졌으며 음변화가 있는 단어에 비해 음변화가 없는 단어에서 활성화 수준이 더 떨어졌다.

fMRI를 이용한 실험 결과는 어느 정도 반응시간 자료와 일치한다. 반응시간 실험 결과에서도 저빈도 단어인

경우에 특히 자연음과 합성음간의 차이가 있었다. fMRI 결과에서는 저빈도인 경우에 청각 담당 영역의 저조한 활성화로 나타났다. 특히 저빈도이면서 규칙단어인 경우에 청각 영역에서의 활성화가 낮았는데 이런 결과는 반응 시간 자료와 일맥상통하는 것이다.

IV. 종합 논의

본 연구에서는 자연음 어휘를 정보처리할 때와 합성음 어휘를 정보처리할 때 어떤 차이가 있는지를 조사하고 합성음의 품질에 따른 차이가 무엇인지를 조사하기 위해 실시되었다.

실험 결과에 따르면, 어휘 판단 시간에서는 세 종류의 음성에 따른 차이는 나타나지 않았으며, 빈도효과에서도 큰 차이가 없었다. 그러나 단어규칙성 효과에서는 자연음 조건과 합성음 조건 간에 질적인 차이가 나타났다. 즉, 자연음 조건에서는 단어 규칙성 효과가 나타나지 않았지만 합성음 조건에서는 단어규칙성 효과가 나타났다. 또한 fMRI를 이용한 실험에서도 규칙성이 문제가 되었다. 즉, 단어 빈도와 관련해서는 자연음을 들을 때나 합성음을 들을 때나 큰 차이가 없었지만 단어 규칙성에서는 단어 빈도와 상호작용하여 차이가 있었다. 특히 저빈도이면서 규칙 단어인 경우에 자연음보다 합성음 조건에서 청각 영역의 활성화가 작았다.

따라서, 단어규칙성 효과만을 고려하면 자연음과 합성음 간에는 질적인 차이가 있는 것으로 보인다. 음성 어휘를 듣고 일차적으로 단어인지 아닌지를 판단하는데에는 자연음이나 합성음이나 큰 차이가 없지만 어휘의 속성 등의 심층적인 정보를 처리하는 데에서는 차이가 자연음과 합성음간의 차이가 있는 것이 아닌가 추측된다. 최근에 발표된 Delogu, Conte, Sementina(1998)의 결과를 보면, 문장 이해 시에 합성음이 자연음에 비해 심적 부담(mental load)이 더 커서 더 어렵고, 합성음을 청취할 때 더 큰 주의(attention)가 요구된다고 한다. 이런 맥락에서 본 연구의 결과를 생각해 볼 때 합성음 어휘 정보처리 시에 단어 규칙성과 같은 어휘의 심층적 특성을 다루기 위해서는 더 큰 인지적 부담(cognitive load)이 요구되는 것으로 해석할 수 있겠다.

재미있는 현상은 어휘 판단 시간과 단어 규칙성 효과에서 고품질 합성음과 저품질 합성음 간에는 차이가 없다는 것이다. 표 1에서 볼 수 있듯이 어휘 판단 시간이 두 조건 간에 유사하고 단어 규칙성 효과에서도 유사하다는 것이다. 이런 결과를 놓고 볼 때, 합성음을 어느 정도 식별할 수만 있으면 어휘를 표면적인 수준에서 이해하는 데에는 큰 차이가 없는 것으로 생각된다. 이런 특성은 음성 정보의 애매모호함(ambiguity) 때문에 나타나는 현상이 아닌가 추측된다. 잘 알려진 것처럼 음성 정보는 상당히 애매모호하다. 이런 이유로 때로는 적절한 문맥이 없으면 무슨 단어인지를 이해하기 힘든 경우가 많다. 본 연구에서는 문맥 속에 어휘가 제시된 것이 아니고 음성어휘만이 제시되었기 때문에 음성 어휘의 애매모호함이 그 대로 유지되었을 것으로 예측된다. 이런 애매모호함이 큰 상황에서 일부 정교한 음성 정보가 더 주어지더라도 심층적인 정보를 이해하는 데에는 별 도움이 안 된다면, 고품질 합성음이 저품질 합성음에 비해 더 많이 제공하는 정보가 별 영향력을 가지지 못할 것이다.

본 연구 결과를 종합해 볼 때, 합성음은 자연음에 비해 어휘 자체를 이해하는 과정에서는 큰 차이가 없지만 좀 더 인지적이고 상위의 정보처리가 요구되는 과정에는 어려움을 주는 것으로 볼 수 있겠다.

참고문헌

이광오 . "한글 글자열의 음독과 음운 규칙." *한국심리학회지: 실험 및 인지*, 8, 1., 1996

D. A, Balota,. Visual word recognition: The journey from features to meaning. In Morton ann Gernsbacher (Eds.), *Handbook of Psycholinguistics*: Academic Press.1994

C. Delogu, S. Conte, & C. Sementina, "Cognitive factors in the evaluation of synthetic speech." *Speech communication,* 24, 1994

M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listeners." *Speech communication*, 16, 1995

D. Pisoni, & S. Hunnicutt, "Perceptual evaluation of MITalk: The MIT unresticted text-to-speech system." *Proceedings of ICASSP*, 80, 3, 1980