한국어 음성합성기 성능평가에 의한 합성 음질개선

양희식,*한민수,**김종진

한국정보통신대학원대학교,**전자통신연구원

Speech Quality Improvement by Speech Quality Evaluation

Hee-Sik Yang,*Minsoo Hahn,**Jong-Jin Kim

Information and Communications University,ETRI

sheik@icu.ac.kr

<본 연구는 전자통신연구원(ETRI)의 연구비 지원에 의해서 이루어졌습니다.>

ABSTRACT

본 논문에서는 한국어 합성기의 명료도 및 자연성 평가방안에 대한 개략적인 설명과 이 방안을 실제로 2종류의 서로 다른 한국어 합성기에 적용한 결과를 요약하였다. 한편, 이러한 평가결과를 바탕으로 실제로 이루어진 음질 개선 실 예를 소개하는 한편 향후 한국어 합성기의 성능 개선 방향을 제안하였다.

1.서론

음성 합성 기술은 합성음의 품질이 어느 정도 보장될 경우 그 활용 분야가 무척 다양하다. 즉 컴퓨터와 인간 간의 음성 인터페이스의 구현에 직접 이용될 수 있을 뿐만 아니라 발성 장애자의 대리 발성 도구, 전자 메일 음성 낭독기, 음성 자동안내 시스템 등에 사용될 수 있다. 따라서 이러한 유용성과 다양성 때문에 미국, 유럽, 일본 뿐 만 아니라 우리나라에서도 많은 연구소와 업체 및 대학에서 지속적으로 연구되어 왔으며 지난 20 ~ 30년 간 합성음질 측면에서 괄목할 만한 발전을 이루어 왔다. 특히 반도체 기술의 빠른 발전으로 보다 빠른 CPU 속도와 초대용량 메모리의 사용이 가능해 짐에 따라 대용량 음성 데이터베이스의 활용과 합성 알고리즘의 실시간 처리가 가능해짐에 따라 보다 좋은 음질의 합성음을 보다 저렴한 비용으로 생성할 수 있게 되었다. 각 연구소와 기업 및 대학에서는 독자적인 알고리즘 및 시스템을 이용하여 합성 시스템을 개발하고 개선하였으며 이에 따라서 현재 합성기의 음질은 10년 전에 비해 월등하게 명료하고 자연스러워져 실제 생활의 많은 분야에 거부감 없이 적용할 수 있는 정도의 수준에 도달하였다. 그러나 합성음의 음질이 얼마나 향상되었는지는 음성의 특성상 정량적으로 표현하는 것이 거의 불가능하며 대부분이 청취자의 인지에 의한 주관적 평가로 이루어진다[1]. 따라서 현재 만족할 만한 합성음의 객관적 품질 평가 기준은 확립되어 있지 못하며 대부분의 평가 방법이 MOS(Mean Opinion Score)에 크게 의존하고 있다. 그러나 이 방법은 여러 가지 조건, 예를 들면 피험자의 개인적 취향 및 선호하는 음색과 성장 배경 (특히 언어학적 의미의), 청취 평가 실험이 이루어지는 공간과 합성음 재생 도구의 특성 등이 포함된 실험환경, 평가대상 문장을

이루는 단어 등의 음성학적 특성 등에 의해서 그 점수가 크게 영향을 받는 등 객관성이 결여되어 있다. 즉, 평가 문장, 피험자의 선정, 평가 공간 및 합성음 재생 도구의 특성 등에 대한 객관화가 없이 이루어진 평가 결과는 그 신뢰성이 떨어질 수 밖에 없으므로 이들에 대한 객관화가 필요한 것이다.

음성합성의 전문가가 아닌 사용자의 입장에서도 여러 가지 합성기 중 최적의 합성기를 선택하기 위해선 보다 객관적인 평가 방법 및 절차에 의한 평가 결과가 필요하며, 개발자의 입장에서는 합성기 내에서 어떤 부분에 문제가 있는지를 인지하고 합성 시스템의 성능을 개선하기 위한 정보를 얻기 위하여 좀더 세부적인 항목까지 포함하는 평가 방법이 요구된다.

본 연구에서는 두 가지 요구, 즉 사용자 입장에서의 객관적 합성음 품질 평가 기준에 대한 요구 및 개발자 입장에서의, 즉 기존의 합성기의 성능 개선에 필요한 정보를 얻기 위한 보다 상세한 음질 평가에 대한 요구를 충족시키기 위한 한국어 합성기 평가 방안을 제시하여 현재 여러 연구소 및 기업에서 개발되고 있는 한국어 합성기들의 평가에 공통적으로 사용될 수 있게 하고자 노력하였으며 향후 이 결과를 바탕으로 적법한 절차에 의한 합성기 성능 평가 국가표준안이 도출되어 향후 국내 음성 관련 시장의 활성화에 기여할 수 있기를 기대한다.

2.본론

2.1.평가 항목의 선정 및 실험 절차의 개발

합성 음질의 평가는 평가 문장, 피험자의 선정, 평가 공간, 실험환경, 실험절차 및 실험 도구 등에 의해 동일한 시스템에 대해서도 그 평가 결과는 변화할 수 있다는 것을 이미 언급하였다[1]. 따라서 이러한 변수들에 대한 객관화가 필수적이며 또한 개별 평가 항목에 대한 공통적인 기준 또한 필요하다. 즉 서로 다른 합성 시스템에 대해 동일한 잣대를 가지고 평가하고 실험 변수 또한 오차범위 내에서 동일하게 적용된다면 그 평가 결과는 사용자 혹은 개발자에게 서로 다른 시스템을 비교하기 위한 유용한 수치가 될 것이다.

동일한 평가 기준 및 오차 범위 내의 실험 변수들을 가지도록 하기 위하여 먼저 최근 한국어 합성 시스템의 기술 동향을 반영한 평가 항목들을 선정하였다. 평가 항목들은 크게 명료도, 자연성, 이해도, 선호도, 자유발화정도, ERP스트레스강도, 청취피로도, 잡음환경 민감도, 발음변환 정확도, 형태소분석 정확도 및 끊어 읽기 정확도로 구성되며 각각의 항목들은 항목별 특성에 따라 세부 항목으로 나누어 평가 할 수 있도록 하였다. 이 중 합성음질 평가 결과에 중요도가 높은 명료도, 이해도, 자연성 및 선호도를 필수적으로 평가하게 하였고 나머지 항목은 선택적으로 평가 할 수 있도록 하였다. 본 연구에서는 필수 항목들에 대한 의미의 정의, 항목별 세부 항목 및 평가 대상 음성의 선정을 우선적으로 진행하고 선택항목에 대한 평가 방법은 차후의 과제로 남겨두었다. 평가의 방법은 듣기평가 방법을 채택하였고 합성음질평가의 실험 변수들이 편차가 유효한 오차범위 내에 있도록 하기 위하여 [2] 및 ITU-T Recommendation[3][4][5][6]에 기초하여 실험 절차, 실험 방법 및 실험 환경을 규정 하였으며 개략적인 실험의 절차는 그림2.1과 같다.

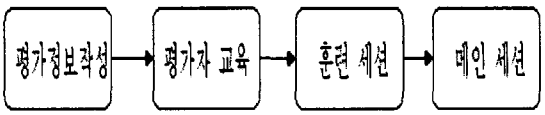


그림2.1 합성음 듣기 평가의 절차

2.2.평가 항목 및 실험 절차 검증을 위한 합성 시스템 평가

실험 절차 및 선정 항목에 대한 유효성 검증을 위하여 국내 회사에서 개발된 두 개의 합성시스템 A,B를 대상으로 합성 음질 평가 필수 항목 중 이해도 및 자연성 평가를 실시하였다.

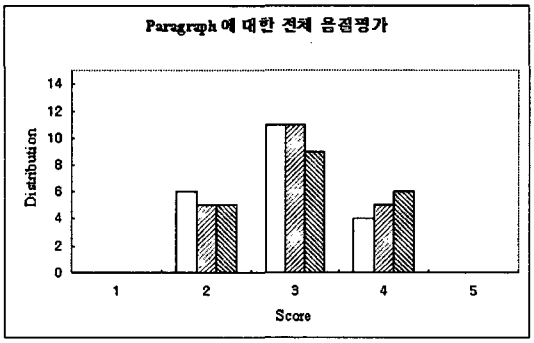
2.2.1.실험 변수

청취 평가에 적용되는 변수들이 객관성을 가지도록 평가 문장, 피험자의 선정, 평가 공간, 실험 환경, 실험

절차 및 실험 도구 등에 대한 설정은 사전에 정해진 규정에 따랐으며 그 상세 내역은 다음과 같다. 실험 환경은 15m^{2}의 사무실 환경 잡음 환경에서 Loud Speaker System을 청취시스템으로 사용하였으며 청취 레벨은 30~40dB 사이에서 미세 조정하며 피험자가 선정토록 하였다. 평가 대상음성은 신문 잡지 및 인터넷에서 추출한 10~20초 시간 분량의 메시지와 수필에서 추출한 1분30초~ 2분 시간 분량의 구문을 이용하였으며 피험자는 ITU-T Recommendation P.85의 기준에 따라 선정하고 합성음질 테스트를 실시하였다[3].

2.2.2.합성 시스템 평가

두개의 합성 시스템에 대하여 메시지에 대한 MOS, 구문에 대한 MOS를 측정하였으며 상세 항목으로 이해도, 청취 노력도, 발음품질, 발화 속도 및 청취 선호도를 측정하였다. 구문 평가의 경우 시스템 A,B가 각각 이해도 2.90, 3.48, 청취노력도 3.05, 3.65, 발음품질 3.05, 3.10, 발화속도 3.71, 3.35(발화속도는 3이 적절한 속도라고 설정함), 청취 선호도 2.52, 2.75의 결과를 보였으며 전체 음질은 각각 2.90 및 3.05의 결과를 보였다. 메시지를 평가음성으로 사용한 경우 시스템 A,B의 평가 결과는 각각 이해도 2.73, 3.53, 청취노력도 2.74, 3.79, 발음품질 2.66, 3.34, 발화속도 3.98, 3.63(발화속도는 3이 적절한 속도라고 설정함), 청취 선호도 2.40, 2.97의 결과를 보였으며 전체 음질은 각각 2.72 및 3.44의 결과를 보였다. 시스템 A, B에 대한 평가 결과는 각 시스템의 개발자들에게 피드백 되었으며 시스템 A의 경우 평가 결과를 이용하여 시스템의 개선이 이루어졌다. 개선된 시스템 A의 경우 개선 전 시스템과 비교할 때 구문 평가에서는 이해도 19.67%, 청취노력도 28.13%, 발음품질 4.69%, 발화속도 15.38%, 청취 선호도7.55% 개선되어 전체 음질이 MOS 2.9에서 3.0으로 3.3% 향상되었으며, 메시지 평가에서는 이해도 31.17%, 청취노력도 37.87%, 발음품질 30.23%, 발화속도 17.49%, 청취 선호도 22.38% 개선되어 MOS 2.72에서 3.26으로 19%의 전체 음질 향상이 있었다. 2개의 합성 시스템 A,B의 평가 결과 및 개선된 합성시스템 A의 합성시스템의 MOS 평가 결과를 그림2.2에 도시하였다. 왼쪽부터 시스템 A, 시스템 A의 개선된 버전, 시스템 B의 평가 결과 분포를 표시하였다.



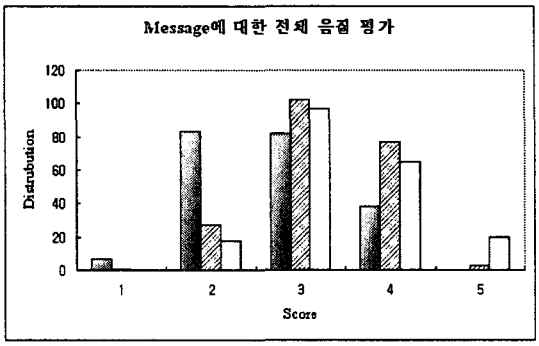


그림2.2 시스템 A,B에 대한 MOS 분포

3.결론

이상 기술한 바와 같이 본 연구에서는 합성 음질 평가에서 가장 중요한 듣기 평가 절차의 1차 버전이 완성되었고 이를 이용하여 잘 알려진 두 종류의 한국어 합성기에 대한 품질 평가를 수행하였다. 강조하고 싶은 점은 합성기 품질 평가 결과를 합성기 개발자에게 피드백하고 제시된 문제점에 대응하여 약간의 노력을 경주한 결과 개선된 합성기의 품질이 눈에 띌 만큼 향상되었다는 것이다. 이는 규격화된 합성 음질 평가 방안이 한국어 합성기의 전반적인 품질 향상을 위해서 반드시 필요하다는 것을 입증하는 실례라 하겠다.

향후 해야 할 일은 아직 명확히 정의되지 않은 부수적인 세부 항목들에 대한 의미 및 정의를 마무리하고 그에 따른 절차를 만들어 가는 것과 이미 정의된 평가 항목에 대한 심도있는 토론을 거친 수정일 것이다. 이러한 과정을 거친 후 적합한 절차에 따라 국가

표준화 함으로써 음성 관련 기술의 투명성 재고 및 시장 활성화에 기여할 수 있을 것으로 기대된다.

참고문헌

[1] 오영환. 1998. 음성언어 정보 처리 : 음성합성, pp 211-239.

[2] Sebastian Möller, 2000, Assessment and prediction of speech quality in telecommunications, pp 121-129.

[3] ITU-T Recommendation P.85, 1994, "A method for subjective performance assessment of the quality of speech voice output devices".

[4] ITU-T Recommendation P.800, 1993, "Method for subjective determination of transmission quality".

[5] ITU-T Recommendation P.64, 1993, "Determination of sensitivity/frequency characteristics of local telephone systems".

[6] ITU-T Recommendation P.810, 1996, "Modulated Noise Reference Unit(MNRU)".