말소리와 음성과학 제2권 제1호 (2010)
pp. 71~75

HMM 기반의 한국어 합성음에 대한 PESQ 및 MOS 평가의 상관도 분석

Correlation Analysis of PESQ and MOS Evaluation for HMM-based Synthetic Korean Speech

임 창 송[1] · 배 건 성[2]
Lin, Cangsong* · Bae, Keunsung**

ABSTRACT

The PESQ is an objective speech quality evaluation measure that is known to have a high correlation with a subjective speech quality measure such as MOS. To examine whether it could be useful as an objective quality measure of synthetic speech, we carried out both subjective evaluation tests with MOS and DMOS and an objective evaluation test with PESQ for HMM-based Korean synthetic speech signals and analyzed the correlation between them. Experimental results have shown that the PESQ has correlations of 0.87 with MOS and 0.92 with DMOS. It means that the PESQ holds much promise for evaluating the quality of synthetic Korean speech.

Keywords: HTS, PESQ, MOS, speech quality evaluation

1. 서론

음성합성 기술은 오늘날 낭독시스템, 음성 자동안내시스템, 전자책 등과 같이 다양한 분야에서 사용되고 있다. 이에 따라 보다 자연스럽고 명료한 합성음을 만들기 위한 연구가 꾸준히 진행되면서 신뢰성 있는 합성음의 음질평가 방법에 대한 연구 필요성도 대두되고 있다[1,2]. 음성신호에 대한 음질평가 방법에 대한 연구는 음성통신 분야에서 오랫동안 연구되어 왔는데, 음성의 특성을 객관적으로 정량화하여 표현하는 것이 어렵기 때문에 대부분 청취자의 인지에 의한 주관적 평가로 이루어진다. 그러나 음질의 주관적 평가는 피험자의 개인적 취향 및 선호하는 음색, 청취 평가 실험이 이루어지는 환경, 평가대상 문장의 음성학적 특성 등에 의해 그 결과가 크게 영향을 받게 된다. 특히, 주관적 음질평가 방법은 절차는 간단하지만 그 과정을 준비하고 실행하는데 많은 시간과 비용이 소요된다는 단점이 있다. 따라서 주관적 음질평가를 대체할 수 있는 객관적 음질평가에 대한 연구가 꾸준히 진행되면서 ITU-T를 중심으로 표준안이 발표되어 왔는데, 가장 최근에 PESQ(Perceptual Evaluation of Speech Quality) 방식이 표준안으로 권고되었다[3].

최근에 주로 많이 이용되는 음성합성 기술은 음소나 음절 등의 단위로 음성신호의 기본 파형을 추출하여 저장하고 이를 이용하여 합성하는 코퍼스(corpus) 기반의 음성합성 방식이다. 코퍼스 기반의 음성합성 방식은 높은 음질의 합성음을 만들어 내지만 대용량의 음성 DB를 필요로 하고, 음성 DB의 레이블 정보 등이 필요하므로 DB 구축에 많은 시간과 비용을 필요로 한다. 또한 대용량의 음성 DB로 인해 메모리 크기에 제한 받는 임베디드 시스템에서는 구현이 어렵다는 단점이 있다. 이에 반해 최근에 많이 연구되고 있는 HMM(Hidden Markov Model) 기반의 음성합성 방식은 합성에 필요한 파라미터를 통계적인 음향모델로 표현하므로 전체 시스템의 메모리 사용이 적으면서도 양호한 음질의 합성음을 얻을 수 있어 임베디드 시스템에의 적용이 용이하고, 적은 음성 DB로도 쉽게 합성음시스템을 구현할 수 있다는 장점이 있다. 이러한 HMM 기반의 음성합성 방식에 대한 연구가 일본 및 유럽에서 활발하게 진행되면서 HTS (HMM-based Text-to-Speech System)의 소스도 공개되어[4] 이를 이용한 한국어 합성에 대한 연구도 진행되고 있다[5-7].

본 논문에서는 음성통신 분야에서 주관적 음질평가 결과인 MOS(Mean Opinion Score)와 가장 높은 상관도를 갖는 객관적

1) 경북대학교 truker@lycos.co.kr
2) 경북대학교 ksbae@ee.knu.ac.kr, 교신저자

접수일자: 2010년 1월 27일
수정일자: 2010년 3월 9일
게재결정: 2010년 3월 14일

음질평가 방법으로 ITU-T 표준안으로 제정된 PESQ 척도를 한국어 합성음의 객관적 음질평가 방법으로 사용할 수 있는지의 타당성을 검증하고자 한다. 이를 위해 HTS를 이용하여 생성한 15개 문장의 한국어 합성음에 대해 주관적 음질평가 실험을 수행하고, 그 결과인 MOS 및 DMOS(Degradation MOS)와 객관적 음질평가 방법인 PESQ를 적용하여 얻은 득점과의 상관도를 분석하였다. 실험에 사용된 전체 합성음에 대한 분석에서 MOS와 PESQ, DMOS와 PESQ 득점 사이의 상관계수는 각각 0.87 및 0.92로 높은 상관도를 얻음으로써 PESQ를 합성음의 객관적 음질평가 척도로 사용할 수 있음을 보였다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 음질평가 방법에 대해 서술하고 객관적 음질평가 방법인 PESQ 척도(measure)에 대해 간단히 설명한다. 3장에서는 합성음의 주관적 음질평가와 객관적 음질평가에 대한 실험 내용과 결과를 제시하며, 마지막으로 4장에서 결론을 맺는다.

2. 음질평가 방법

음성의 음질평가 방법은 <그림 1>과 같이 일반적으로 통신시스템을 통과하여 생성된 음성을 사람이 듣고 절대적인 평가를 수행하는 주관적 음질평가 방법과 원래 음성과 시스템을 통과하여 생성된 음성을 다양한 척도들을 이용한 수학적 계산을 통해 비교하여 평가하는 객관적 음질평가 방법으로 분류할 수 있다.

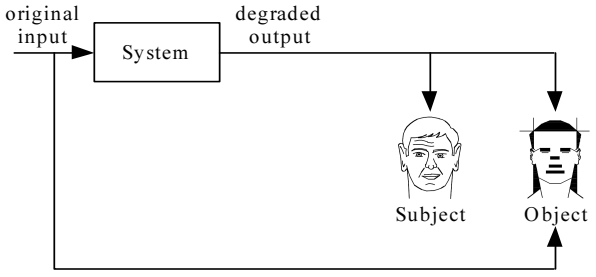

그림 1. 음성의 음질평가 방법
Figure 1. Evaluation of speech quality method

2.1 주관적 음질평가 방법[8]

주관적 음질평가란 실제로 사람이 평가하고자 하는 대상 음성을 듣고 느끼는 심리적 평가 결과에 근거하여 음성품질을 판정하는 것을 말한다. 청취자는 평가 실험에 사용되는 문장의 음성을 듣고 미리 정해진 등급 분류에 따라 등급을 판정하여 점수를 매기고, 평가에 참여한 모든 청취자들의 평가 점수를 평균한 값을 그 결과로 사용하는데 이를 평균청취음질평가점수(mean listening quality opinion score)라고 하며, 간략히 줄여서 MOS라고 한다. <표 1>은 MOS 평가에 사용되는 5개의 음질평가 등급과 해당 점수를 보인 것이다.

표 1. MOS 평가 등급
Table 1. MOS scale

Quality of speech	score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

주관적 음질평가 방법은 평가하고자 하는 음성만 듣고 평가하는 절대분류등급(absolute category rating; ACR), 평가하고자 하는 음성이 원래의 기준 음성에 비해 얼마나 열화 되었는지를 평가하는 열화분류등급(degradation category rating, DCR) 등으로 나뉘는데, 앞에서 설명한 MOS는 ACR 방법에 의한 평가 결과를 의미하고, DCR 방법에 의한 결과는 DMOS라고 한다. ITU-T에서는 통신시스템에서의 주관적 평가 방법으로 ACR 방법을 권장하고 있으며, DMOS는 평가하고자 하는 음성이 잡음 음성과 같이 절대적으로 음질이 좋지 못할 때 적합하다. 또한, DMOS는 기준음성과 평가하고자 하는 음성을 동시에 제공하는 비교 평가이기 때문에 음질의 절대적인 평가는 어렵지만 음질을 평가함에 있어서 MOS 보다도 음질 차이를 구분할 수 있는 감도는 높다. 합성음의 경우 동일한 음성에 대해 왜곡된 음성을 평가하는 일반적인 통신시스템과는 달리 기준음성과 의미만 같을 뿐 음운 정보 등이 완전히 다른 새로운 음성이 생성되므로 비교 평가방법이 더 적절할 수 있다. 따라서 본 논문에서는 합성음의 주관적 음질평가 방법으로 MOS 및 DMOS 평가를 수행하였다. <표 2>는 DMOS 평가에 사용되는 5개의 음질평가 등급과 해당 점수를 보인 것이다.

표 2. DMOS 평가 등급
Table 2. DMOS scale

Degradation evaluation	score
Degradation is inaudible	5
Degradation is audible but not annoying	4
Degradation is slightly annoying	3
Degradation is annoying	2
Degradation is very annoying	1

2.2 객관적 음질평가 방법 및 PESQ

주관적 음질평가는 실험환경과 개인적 요인에 따라 차이를 보일 수 있으며, 특히, 그 과정을 준비하고 실행하는데 많은 시간과 비용이 소요되는 단점이 있다. 이러한 단점을 해결하기 위해 주관적 음질평가의 결과와 높은 상관도를 유지하면서 아를 대체할 수 있는 객관적 음질평가 척도를 개발하고자 하는 연구가 꾸준히 진행되어 왔다[3,9,10]. 객관적 음질평가는 원래의 음성을 기준으로 하여 평가하고자 하는 음성과의 차이를 나타내는 척도를 이용하게 되는데, 시간영역에서의 척도로는 신호대잡음비 등이 있고, 주파수영역에서의 척도로는 두 음성 사이의

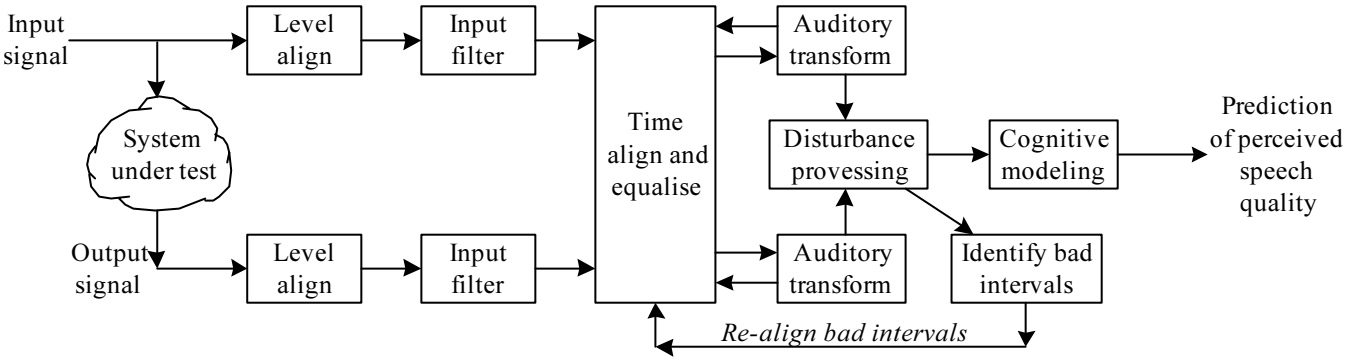

그림 2. PESQ 척도를 구하는 과정의 블록도
Figure 2. Block diagram of the process for PESQ measure

스펙트럼 거리, 켑스트럼 거리 등이 있다. 그러나 단순한 시간 영역 또는 주파수영역에서의 거리로는 객관적 음질평가를 위한 만족스러운 결과를 얻을 수 없고, Bark scale과 같은 사람이 음성을 지각하는 영역으로의 변환을 통해서 거리를 구하는 척도가 주로 이용된다. 전화대역에서의 음성코덱 뿐만 아니라 단말(end-to-end) 사이의 객관적 음질평가를 위한 ITU-T의 P.862 권고안으로 제정된 PESQ는 현재 MOS와 가장 높은 상관도를 갖는 객관적 음질평가 방법으로 알려져 있다.

PESQ는 전화대역의 음성신호를 대상으로 개발된 음성품질 측정 알고리즘으로 음성코덱, 변동적인 지연, 필터링, 패킷이나 셀 손실 및 채널 손실을 가지고 있는 시스템에 적용할 수 있도록 고안되었다. PESQ의 기본 개념은 <그림 2>와 같다. 입력 음성신호와 시스템을 통과한 출력 음성신호를 주관적 음질평가에서 일반적으로 사용하는 레벨과 비슷하게 하기 위해서 표준청각 레벨을 기준으로 입력신호의 레벨을 정렬하는 과정과 수화기의 대역통과 특성을 고려하기 위한 필터링 과정을 거치게 된다. 다음으로 왜곡된 음성신호의 지연을 고려하여 두 음성신호의 시작점을 찾아 시간정렬을 수행하고 비가청음 영역의 신호를 제거하는 과정을 거친다. 그리고 원래 음성신호와 왜곡된 음성신호의 소리의 강도(loudness density) 차이를 계산하여 소리의 강도가 낮은 부분에 대해서 다시 시간정렬 과정을 실행하여 보다 정확한 시간정렬 과정을 수행한 후, 사람의 청각특성을 고려한 지각영역으로의 변환과 인지과정을 모델을 거침으로써 최종적으로 -0.5에서 4.5의 범위를 가지는 PESQ 득점을 산출하게 된다.

3. 실험 및 검토

3.1 실험 환경 및 조건

주관적 및 객관적 음질평가 실험을 위해 HTS 2.0.1 버전으로 구현한 한국어 합성시스템[7,11]을 이용하여 15개의 합성음을 만들었다. 사용한 한국어 합성시스템의 HMM 음향모델은 묵음모델을 포함한 47개의 유사음소 음향모델을 초기모델로 하여 임베디드 훈련과 클러스터링 과정을 통해 문맥종속 모델로 확장하였다. 합성기의 음향모델 생성에 사용된 음성 DB는 국립국어원[12]에서 배포한 “서울말 낭독체 발화 말뭉치”인데, 음향모델 훈련에 사용된 646 문장의 음성 중에 길이가 짧은 것, 중간 것, 긴 것 각각 5개씩, 총 15개를 선정하여 기준음성으로 사용하고, 그에 해당되는 각 문장을 입력으로 하여 8㎑ 샘플링 주파수를 갖는 15개의 합성음을 생성하였다. 그런 다음, 생성한 각 합성음에 대해 MOS 및 DMOS 평가를 수행하고, 각각의 결과와 PESQ 득점과의 상관도를 분석하였다. 그림 3은 본 논문에서 수행한 전체 실험과정의 블럭도를 보인 것이다.

주관적 음질평가 실험은 사무실 환경에서 헤드폰을 청취도구로 사용하여 피험자가 듣기 편한 음량 크기로 설정하였다. 주관적 음질평가 척도로는 MOS 척도에 의한 자연성 평가와 DMOS 척도에 의한 기준음성과의 자연성 비교 평가를 하였다. 음성평가 경험이 없는 20대 남성 20명과 20대 여성 20명이 주관적 음질평가에 참여하였는데, MOS 테스트와 DMOS 테스트에는 각각 서로 다른 남성 10명과 여성 10명이 참여하였다. MOS 테스트는 길이가 다른 15개 문장에 대한 합성음을 랜덤한 순서로 청취하면서 표 1의 등급에 따라 판정하고, 그 결과를 합산하여 평균을 구하였다. 기준음성과 평가음성 모두를 들려주는 DMOS 테스트도 같은 방법으로 수행하였다. 객관적 음질평가에는 ITU-T에서 공개하고 있는 PESQ 2.0 버전을 사용하였다[13]. PESQ 척도를 이용한 객관적 음질평가에서는 기준음성과 평가음성을 필요로 하는데, DMOS 테스트에 사용된 15개의 기준음성과 평가음성을 이용하여 PESQ 득점을 산출하였다.

3.2 실험 결과

본 논문에서 수행한 MOS 테스트의 결과는 <그림 3>과 같다. 본 논문에서 실험에 사용된 합성음에 대한 MOS 득점의 평균값은 2.4에서 3.8의 분포를 보이고 있으며, 표준편차는 0.59에서 0.95의 크기를 나타내었다. <그림 4>는 DMOS 테스트의 결과이다. DMOS 득점의 평균값도 MOS와 비슷하게 2.35에서 3.85의 분포를 보이고 표준편차는 0.44에서 0.95의 크기를 나타내고 있다.

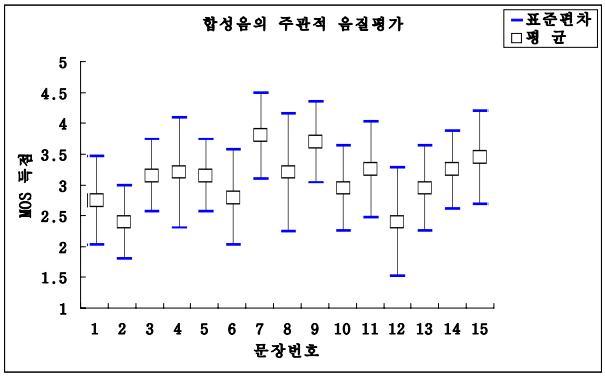

그림 3. 전체 피험자들의 MOS 득점
Figure 3. MOS values for all listeners

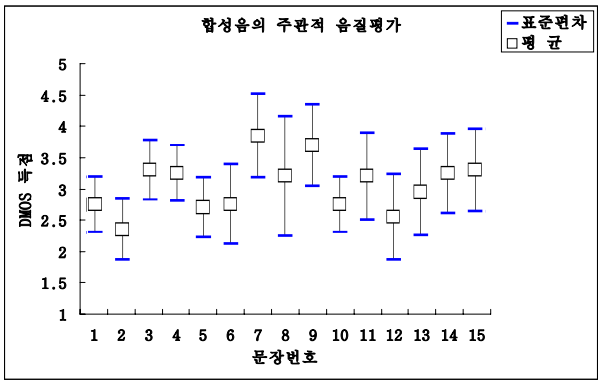

그림 4. 전체 피험자들의 DMOS 득점
Figure 4. DMOS values for all listeners

합성음에 대한 객관적 음질평가 방법으로PESQ 척도의 타당성을 검증하기 위하여 식 (1)에 주어진 Pearson의 상관관계식을 이용하여 주관적 음질평가 결과와 PESQ 득점 사이의 상관도를 조사하였다.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \qquad (1)$$

여기에서, x_i 는 i 번째 문장의 주관적 음질평가 득점, $\bar{x}$ 는 주관적 음질평가 득점의 평균이며, y_i 는 i 번째 문장의 PESQ 득점, $\bar{y}$ 는 PESQ 득점의 평균이다.

<표 3>에 남성 및 여성 평가자 각각에 대한 결과와 전체 평가자에 대한 결과를 나타내었다. <표 3>에서 보면, MOS 평가의 경우 여성 평가자들이 남성 평가자들에 비해 더 높은 상관도를 보였으며, DMOS 평가의 경우에는 남성, 여성 평가자 모두 MOS 보다 약간 높은 상관도를 보였다. 각 문장별로 남성, 여성을 구분하지 않고 전체 평가자에 대해서 구한 PESQ와 MOS 득점 사이의 상관도는 0.87, DMOS와는 0.92로 두 경우 모두 높은 상관도를 나타내었다. 이것은 각 문장에 대한 평가자 수가 2배로 늘어나면서 남성 및 여성평가자 각각의 경우에 비해 평가결과의 변이가 줄어들었기 때문이라고 생각된다.

표 3. 주관적 음질평가 결과와 PESQ 득점 사이의 상관도
Table 3. Correlation between PESQ and MOS, PESQ and DMOS values

관 계 / 피험자	MOS 득점과 PESQ 득점 사이 상관관계 r	DMOS 득점과 PESQ 득점 사이 상관관계 r
남성	0.7538	0.8696
여성	0.8637	0.8885
전체	0.8704	0.9190

<그림 5>과 <그림 6>은 각 평가음성에 대해 전체 평가자들에 대한 MOS와 PESQ, DMOS와 PESQ 득점 사이의 관계를 선형 회귀로 계산한 추세선을 보인 것으로, 점선은 95% 신뢰도를 가지는 예측구간을 나타낸 것이다. 그림에서 보면, MOS 및 DMOS 모두 PESQ와 높은 상관관계를 가지고 있으며, MOS에 비해 DMOS가 약간 더 높은 신뢰성을 가짐을 볼 수 있다.

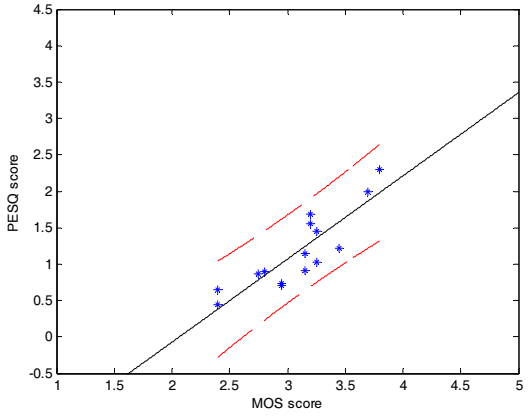

그림 5. 전체 피험자들에 의한 MOS 득점과 PESQ 득점 사이 관계
Figure 5. Relation between MOS and PESQ values

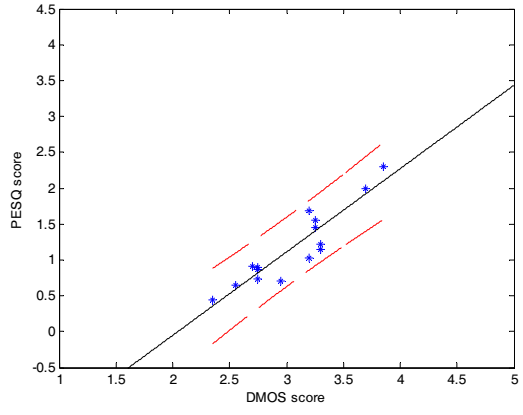

그림 6. 전체 피험자들에 의한 DMOS 득점과 PESQ 득점 사이 관계
Figure 6. Relation between DMOS and PESQ values

4. 결론

본 연구에서는 한국어 합성음의 객관적 음질평가 방법으로 PESQ 척도의 타당성을 알아보기 위한 목적으로 HTS 기반의 합성음에 대한 주관적 평가 실험을 수행하고, 그 결과인 MOS 및 DMOS와 객관적 평가방법인 PESQ를 이용한 득점과의 상관도를 분석하였다. 한국어 합성음에 대해 PESQ를 이용한 객관적 음질평가 결과는 주관적 음질평가 결과인 MOS와는 0.87의 상관도를 보였으며, DMOS와는 0.92정도의 높은 상관도를 보여 한국어 합성음의 객관적 음질평가 방법으로 PESQ 척도가 충분히 사용될 수 있음을 보였다.

향후 보다 다양한 음성과 평가 피험자에 대한 합성음의 주관적 음질평가 결과를 많이 확보하여 신뢰도를 높일 수 있는 방법에 대한 연구가 필요하며, 이를 바탕으로 객관적 음질평가 방법인 PESQ 척도로부터 그에 상응하는 MOS 값을 구할 수 있는 적절한 매핑 함수에 대한 연구가 필요하다.

참고문헌

[1] Cho, C.W., Lee, S.H., Kim, S.J. (2005). “Evaluation guidelines of Korean synthetic speech”, Korea Society of Speech Sciences, Research Report.
(조철우, 이상호, 김수진, (2005). “한국어 합성음평가 가이드라인”, 대한음성학회 연구보고서.)

[2] Yang, H.S., Han, M.S., Kim, J.J. (2002). “Speech quality improvement by speech quality evaluation”, Proceedings of the KSPS Conference, pp. 37-40, Nov.
(양희식, 한민수, 김종진, (2002). “한국어 음성합성기 성능평가에 의한 합성 음질개선”, 대한음성학회 학술대회지, pp. 37-40, 11.)

[3] ITU-T Recommendation (2001). “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”, P.862, 02.

[4] HTS Working Group (2006). “HMM-based Speech Synthesis System”, http://hts.sp.nitech.ac.jp/

[5] Kim, S. J. (2007). “HMM-based Korean speech synthesizer with two-band mixed excitation model for embedded applications”, Ph.D. dissertation, School of Engineering, Information and Communication University, Korea.

[6] Bae, J.C., Bae, K.S. (2007). “HMM-based Korean Speech Synthesis”, Korea Signal Processing Conference, Vol. 20, p. 144.
(배재철, 배건성, (2007). “HMM 기반의 한국어 음성합성”, 신호처리합동학술대회 논문집, Vol. 20, p. 144.)

[7] Kim, I.H. (2008). “Improving naturalness by controlling of duration in HMM-based Korean text-to-speech system”, M.S. thesis, School of Electrical Engineering and Computer Science, Kyungpook National University, Dec.
(김일환, (2008). “HMM 기반 한국어 음성합성에서 지속시간 제어를 통한 자연성 개선”, 석사학위 논문, 경북대학교 대학원 전자전기컴퓨터학부, 12.

[8] ITU-T Recommendation (1996). “Methods for subjective determination of transmission quality”, P.800, 08.

[9] Yang, W., Benbouchta, M., Yantomo, R. (1998). “Performance of the modified Bark spectral Distortion as an objective speech quality measure”, Proceedings of ICASSP, Vol. 1, pp. 541-544.

[10] ITU-T Recommendation (1996). “Objective quality measurement of telephone-band (300-3400Hz)”, P.861, 08.

[11] Kim, I.H., Bae, K.S. (2008). “Control of duration model in HMM-based Korean Sspeech synthesis”, Speech Sciences, Vol. 15, No. 4, pp. 97-105.

[12] http://www.korean.go.kr/

[13] ITU-T Recommendation (2001). “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”, P.862, 02.

• **임창송 (Lin, Cangsong)**
경북대학교 전자전기컴퓨터학부
대구광역시 북구 산격동 1370번지
Tel: 053-940-8627
Email: truker@lycos.co.kr
관심분야: 음성신호처리, 디지털신호처리
2009. 8 전자전기컴퓨터학부 대학원 석사과정 졸업

• **배건성 (Bae, Keunsung)** 교신저자
경북대학교 전자전기컴퓨터학부
대구광역시 북구 산격동 1370번지
Tel: 053-950-5527
Email: ksbae@ee.knu.ac.kr
관심분야: 음성신호처리, 음성합성, 적응신호처리
1979 ~ 현재 경북대학교 전자전기컴퓨터학부 교수