한국ITS학회논문지
제10권, 제4호 (2011년 8월)
pp.71~77

모바일 VoIP 음성통신을 위한 대화음질 측정 시스템

Conversational Quality Measurement System for Mobile VoIP Speech Communication

조 재 만* (Jae-Man Cho)　　김 형 국** (Hyoung-Gook Kim)

요 약

본 논문에서는 고품질 모바일 VoIP 음성통신에 대한 객관적인 QoS를 제공하는 대화음질 측정시스템을 구현하였다. 대화음질 측정을 위해서 VoIP로 연결된 두 대의 스마트폰에 에코 및 잡음 제거, 음성 인코딩 및 디코딩, RTP (Real-Time Protocol)을 적용한 패킷 생성, 지터버퍼 콘트롤, LC (Loss Concealment)를 포함한 POS (Play-out Schedule)로 구성된 VoIP 음성 통화시스템을 구현하였다. 대화음질 측정 시스템은 VoIP로 연결된 두 스마트폰의 마이크, 그리고 스피커와 연결되어 각 화자별로 음성신호를 녹음한 후에, 녹음된 음성신호를 이용하여 CE (Conversational Efficiency), CS (Conversational Symmetry) 및 PESQ (Perceptual Evaluation of Speech Quality)를 측정하고, CE-CS-PESQ에 대한 상관관계를 측정한다. 본 논문에서는 다양한 SNR, IP 네트워크망 변동에 따른 지연, 손실 변화에 따른 CE, CS, PESQ를 측정하여 대화음질 측정시스템을 검증하였다.

Abstract

In this paper, we propose a conversational quality measurement (CQM) system for providing the objective QoS of high quality mobile VoIP voice telecommunication. For measuring the conversational quality, the VoIP telecommunication system is implemented in two smart phones connected with VoIP. The VoIP telecommunication system consists of echo cancellation, noise reduction, speech encoding/decoding, packet generation with RTP (Real-Time Protocol), jitter buffer control and POS (Play-out Schedule) with LC (loss Concealment). The CQM system is connected to a microphone and a speaker of each smart phone. The voice signal of each speaker is recorded and used to measure CE (Conversational Efficiency), CS (Conversational Symmetry), PESQ (Perceptual Evaluation of Speech Quality) and CE-CS-PESQ correlation. We prove the CQM system by measuring CE, CS and PESQ under various SNR, delay and loss due to IP network environment.

Key words : Mobile VoIP QoS, perceptual evaluation of speech quality, conversational efficiency, conversational symmetry

† 이 논문은 2011년도 광운대학교 교내 학술연구비 지원에 의해 연구되었음.
이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2011-0004311)
* 주저자 : 광운대학교 전자융합공학과 석사과정
** 공저자 및 교신저자 : 광운대학교 전자융합공학과
† 논문접수일 : 2011년 1월 10일
† 논문심사일 : 2011년 7월 28일
† 게재확정일 : 2011년 7월 29일

Ⅰ. 서 론

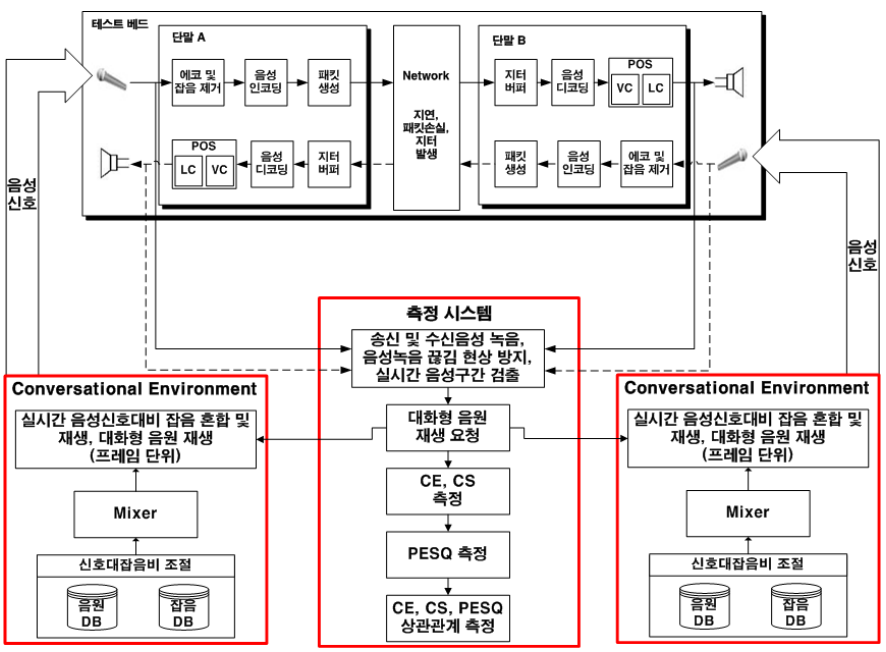
2G, 3G, 3.5G 등의 이동통신 시장을 넘어 많은 사용자들이 사용하고 있는 서비스인 VoIP (Voice over Internet Protocol)는 기존에 제공되던 유선 VoIP를 벗어나 모바일 VoIP 형태로 진화하고 있다. 그러나 무선 IP 네트워크상에서 발생하는 네트워크 과부하, 패킷 지연, 지터와 패킷 손실 등의 문제들이 모바일 VoIP 서비스의 통화품질에 심각한 영향을 미치고 있다. 이러한 모바일 VoIP 통화품질을 개선하기 위해서는 기존의 수신부측에서 수신음성에 대한 PESQ[1], MOS Score[2]를 측정하는 방식을 벗어나서 모바일 VoIP시스템 전체를 구성하는 송신부, 네트워크망 변동과 수신부에 대한 전체적인 상관관계를 분석 및 검증하는 것이 중요하다.

본 논문에서는 모바일 VoIP 음성통신 시스템을 구축하고, 구축된 시스템을 기반으로 음성통화 시에 객관적인 통화 QoS를 제공할 수 있는 대화음질 측정(검증) 시스템을 구현하였다. 구축된 통화음질 측정시스템을 통해 다양한 모바일 VoIP 네트워크 환경에 따라 기존의 PESQ 측정뿐만 아니라 지연 정보를 활용할 수 있는 CE, CS를 측정함으로써 모바일 VoIP 통화품질의 성능을 상세하고 명확하게 평가할 수 있다. 이와 함께 PESQ, CE, CS[3, 4]의 상관관계 분석을 통해 통화품질을 개선시킬 수 있는 보다 나은 방식을 테스트 할 수 있기 때문에 기존의 단순한 단방향 음질 측정 장비에 비해 유용하다.

본 논문의 구성은 다음과 같다. 전체 시스템 구조와 세부 알고리즘에 대해 II 장에서 설명하고, III 장에서는 대화음질 측정 시스템의 실험결과 및 고찰에 대해서 설명한다. 마지막 IV장에서는 결론을 서술한다.

Ⅱ. 전체 시스템 구조

본 논문에서 제안하는 VoIP 대화음질 측정시스템의 구성도는 <그림 1>과 같으며, VoIP로 연결된 두 대의 모바일 VoIP 단말기와 대화 음질을 측정하는 측정시스템, 대화 환경 DB로 구성된다. 단말 내의 VoIP 시스템은 기본적인 에코 및 잡음 제거, 음성 인코딩 및 디



〈그림 1〉 VoIP 대화음질 측정시스템 구성도
〈Fig. 1〉 VoIP conversational quality measurement system

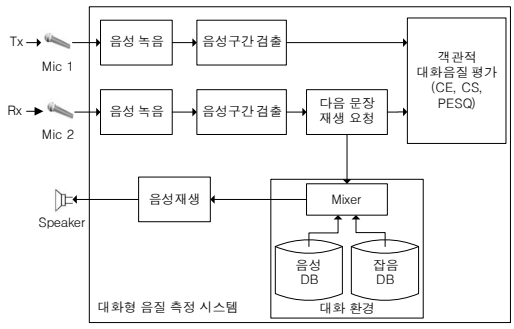
코딩(G.711, G.729, EVRC 등), RTP (Real-Time Protocol)을 적용한 패킷 생성, 지터버퍼, LC (Loss Concealment)와 VC (Voice Classifier)를 포함한 POS (Play-out Schedule)로 구성되며 다음과 같은 과정을 통해 음성통화 및 대화음질 검증이 수행된다.

먼저, 모바일 VoIP 단말기의 송·수신단측 마이크와 연결된 통화환경 (Conversational Environment) 모듈에서 음원DB로부터 통화에 사용되는 음성신호를, 그리고 잡음DB로부터는 거리, 사무실, 공항 등의 환경잡음에 해당하는 잡음신호를 선택한다. 이후에, Signal-to-Noise Ratio (SNR)에 따라 선택된 음성신호에 따른 잡음세기를 선택하게 되면, 음성과 잡음간의 합성이 프레임 단위로 이뤄지면서 배경잡음에 노출된 음성신호가 생성된다. 생성된 음성신호는 단말 A의 마이크와 측정시스템에 동시에 입력된다. 단말 A의 마이크에 입력된 대화 및 통화환경 음성은 에코 및 잡음제거, 음성 인코딩, 패킷 생성을 거쳐 IP 네트워크로 전달된다. 전달되는 패킷은 다양한 지연, 패킷손실, 지터 등을 발생시키는 IP 네트워크를 거쳐 단말 B의 음성 디코딩부로 전달되어 수신된 패킷으로부터 음성데이터를 음성신호로 디코딩하여 지터버퍼, POS을 거쳐 스피커를 통해 음성이 재생된다. 단말의 스피커를 통해 재생되는 음성신호로부터 실시간으로 음성구간을 검출하고 상대방의 음성구간이 종료되었다고 판단되면, 단말 B 측에서도 단말 A의 음성을 듣고 그에 대한 응답 음성으로서 통화 환경에서 생성된 음성신호를 단말 B의 마이크를 통해 동일한 과정을 거쳐 단말 A의 스피커를 통해 상대방에게 전달한다. 통화음질 측정시스템에서는 이 과정 중 발생하는 입력음성과 출력음성을 녹음하고 이를 분석하여 통화 효율성 CE, 통화 대칭성 CS 및 PESQ를 측정하고, CE-CS-PESQ에 대한 상관관계를 측정한다.

1. 측정 시스템 구조

본 과제에서 제안하는 측정시스템의 작동과정은 <그림 2>와 같다.

두 화자의 음성을 동시에 녹음하기 위해 송·수신



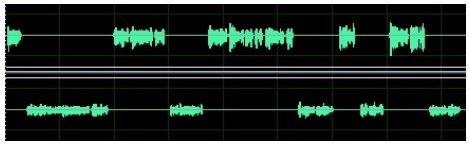
〈그림 2〉 대화음질 측정시스템 작동과정
〈Fig. 2〉 Conversational quality measurement system procedure

음성 녹음은 독립된 사운드카드를 적용하며, 두 화자의 음성은 각각 좌·우 채널에 나뉘어서 녹음된다.

대화 환경기반 음성생성 모듈을 통해 생성된 화자 A의 음성은 마이크에 입력됨과 동시에 송신 음성 녹음부에 바로 입력되어 녹음되며, 수신 음성 녹음부에서는 <그림 1>과 같이 단말을 거쳐 상대 단말의 스피커로부터 재생되는 음성을 입력받아 음성이 녹음된다. 이 때, 음성 구간의 확인을 위해 에너지 기반 음성구간 검출 알고리즘[5]을 적용하여 음성 구간이 끝나는 여부를 판단하게 된다. 본 논문에서는 응답지연시간이 500ms가 발생할 경우 화자의 대화가 종료된 것으로 판단하여 상대 화자의 음성에 대한 재생을 요청한다. 이러한 과정을 거쳐 대화가 종료되면, 객관적 대화음질 평가 모듈에서는 녹음된 입력음성과 출력음성에서 검출된 음성구간 시간정보를 통해 네트워크 지연정보인 Mouth-to-Ear Delay (MED), 응답지연정보인 Human Response Delay (HRD)를 계산하여 CE, CS를 측정하고, 송신 음성과 수신 음성을 비교하여 PESQ를 계산한다.

<그림 3>은 측정 과정을 통해 녹취된 송신, 수신 음성 신호를 나타낸 것이다. <그림 3>의 송신 음성은 각 화자의 음성을 왼쪽, 오른쪽 채널에 나누어서 녹음한 것으로 화자 A가 먼저 대화를 시작하고 화자 B가 응답하는 방식으로 구성된다.

수신음성은 단말로부터 재생되는 음성을 녹취하기 때문에 송신 음성과는 반대 방향부터 녹취가 되는 것을 알 수 있다.



a) 송신 음성(위: 화자 A, 아래: 화자 B)

b) 수신 음성(위: 화자 B, 아래: 화자 A)

〈그림 3〉 녹취된 음성 신호
〈Fig. 3〉 Recorded speech signal

2. CE 측정 방법

CE는 음성통화시에 대화의 효율성을 나타낸다. 즉, 총 통화시간 중에서 지연시간을 제외한 실제로 말하고 듣는 시간을 의미한다. 식 (1)은 이러한 CE의 수식적 표현 방법이다.

$$CE = \frac{Speaking\ time + Heard\ time}{Total\ time\ of\ Call} \quad (1)$$

CE를 측정하기 위해서 식 (1)에서 전체 대화 시간은 말하는 시간과 듣는 시간 외에 지연시간이 포함되며, 지연의 종류에는 MED, HRD 로 구성된다. 화자 A가 말을 하여 화자 B가 화자 A의 음성을 들을 때까지의 지연 ($MED_{A,B}$)을 화자 A가 말하는 시간(A Speak Time)과 화자 B가 듣는 시간 (B Heard Time)을 통해 측정한다. 그리고 화자 B가 화자 A의 음성을 듣고 반응하여 말하는데 까지 걸리는 시간 HRDB를 측정하기 위해 화자 B가 듣는 시간(A Heard Time)와 화자 B가 말하는 시간(B Speak Time)의 차이를 적용한다. IP 네트워크를 통한 음성통신에서는 동일한 대화일지라도 더 긴 MED (Mouth-to-Ear Delay)를 갖는 네트워크에서 요금이 더 발생하게 된다. 그러므로 각 화자는 동일한 대화를 하는데 있어서 높은 CE를 요구하게 된다.

3. CS 측정 방법

CS는 음성통화시에 대화의 비대칭성을 나타낸다. 각 화자는 대화 시에 각기 다른 응답 시간을 갖고 있기 때문에, 마주보고 대화할 때의 아주 작고 균일한 지연과는 다르게 대화의 자연스러움이 낮아진다. 화자 A의 입장에 대한 대칭성을 알기 위해서는 A가 경험하는 최대 MS (Mutual Silence)와 최소 MS의 비를 측정한다. 이러한 CS는 식 (2)와 같이 표현한다.

$$CS_A = \frac{\max(MS_A)}{\min(MS_A)} \quad (2)$$

MS는 실제로 A가 경험하는 묵음구간의 시간 길이를 의미한다. 대화의 형태에 따라 그리고 대화를 주고받는 횟수에 따라 CS는 영향을 받게 된다. 예를 들면, 화자 사이에 대화를 주고받는 횟수가 적은 대화에서는 긴 MED에 의한 CS의 감소는 적게 느껴질 것이다. face-to-face 대화에서는 MS와 HRD는 동일하므로 CS_A와 CS_B는 대략 1이 된다. 그러나 통화 시 지연이 증가하면 묵음 구간은 더 이상 동일할 수 없다. 만약 응답시간이 증가하는 것을 상대방이 인지한다면 상대방도 이에 대응하여 응답을 늦게 함으로써 대화의 품질이 낮아진다. 이러한 CS는 실질적으로 다음과 같이 측정한다.

$$CS_A = \frac{HRD_B + MED_{A,B} + HRD_A}{HRD_A} \quad (3)$$

화자 A의 경우 묵음은 화자 A가 말을 하고 화자 B의 응답을 기다리는 MS_{Aj}와 화자 B의 응답을 듣고 화자 A가 다시 대답하는 MS_{Aj+1}로 구성한다. 묵음 중에서 가장 긴 MS_A와 가장 짧은 MS_A로 CS를 측정하며 일반적으로 가장 큰 MS_A는 $MED_{A,Bj-1}$, HRD_{Bj}, $MED_{B,Aj}$로 구성되고 가장 작은 MS_A는 HRD_{Aj+1}로 구성된다. CS_A는 화자 A측에서 측정된 CS이다. MS는 문장과 문장 사이의 묵음구간을 의미한다. 실제 측정 시에는 다음과 같이 CS가 구성된다.

Ⅲ. 실험 및 결과고찰

1. 실험 방법

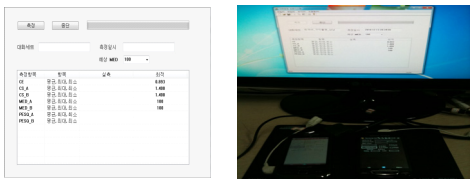
본 논문에서는 구현된 VoIP 대화음질 측정시스템을 검증하기 위한 실험을 다음과 같이 수행하였다.

• 테스트 실험군: 실험을 위해 사용한 음성은 8kHz에서 녹취된 음성이며, 사용된 문장 수는 500문장의 한국어로 구성된다. 총 3회 반복 실험을 하여 SNR (30dB~15dB), 지연변화 (0ms~40ms), 손실변화 (0% ~30%) 에 따른 검증을 실시하였다.

• 테스트 베드에 적용된 VoIP 소프트웨어: 본 논문에 적용된 VoIP 소프트웨어는 본 연구실에서 자체적으로 개발된 소프트웨어로서, 에코 및 잡음제거, VC, 지터추정, 지터버퍼 컨트롤, 음성인코더 및 디코더, 패킷손실 은닉 (LC) 및 병합 알고리즘을 통한 플레이 아웃 스케줄링 (POS) 등으로 구성되었다. 에코 및 잡음제거[6]는 cascaded cross spectral 추정 방식과 LSA (log spectral amplitude) 음성 추정 방식을 적용하였으며, 음성 인코딩 및 디코딩 방식은 G.711을 사용하였다. 지터버퍼 콘트롤 및 POS는 본 연구를 위해 개발된 알고리즘을 적용하였다.

• 모바일 단말기의 환경: 본 실험은 CPU(Core i5 2.8GHz), RAM 4GB의 사양을 갖춘 Windows 7 운영체제에서 CPU(S3C6410 800Mhz), RAM 256MB의 사양을 갖춘 삼성 SPH-M8400모델의 단말 두 대를 이용하여 VoIP 대화음질 측정시스템을 수행하였다.

<그림 4>는 실제 대화음질 측정 화면으로서, 측정이 수행된 후 각 파라미터에 대한 결과가 나타난다.



〈그림 4〉 대화음질 측정시스템 데모
〈Fig. 4〉 Conversational quality measurement system demo

2. 실험결과

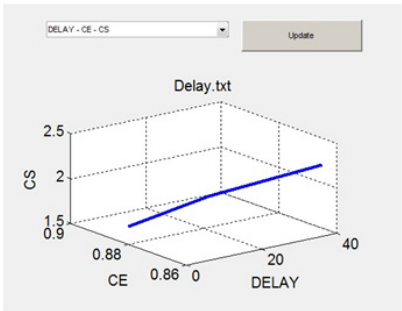
대화음질 검증은 IP 네트워크 망변동에 따른 다양한 지연, SNR, 손실 변화에 따라 각각 측정하였다.

<표 1>은 지연 변화에 따른 CE, CS, PESQ 측정결과를 나타낸다.

〈표 1〉 다양한 IP 네트워크 지연에 따른 실험결과
〈Table 1〉 Experimental results due to various IP network delay

지연	0ms	20ms	40ms
CE	0.880	0.878	0.865
CS	1.700	1.882	2.199
PESQ	4.091	3.994	3.847

<표 1>에 따르면 지연에 의해서 전체 통화시간이 증가하여 CE가 감소하는 것을 알 수 있다. 또한, CS를 구성하고 있는 요소 중 최대 묵음구간이 증가하면서 CS가 증가하는 것을 알 수 있다. 이러한 실험결과는 3차원 그래프를 통해 관찰할 수 있다. <그림 5>에서는 지연에 CE-CS-PESQ의 상관관계를 빠르고 명확하게 판단할 수 있다.



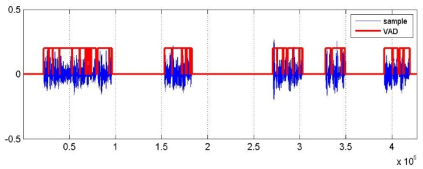
〈그림 5〉 지연에 대한 3D 그래프
〈Fig. 5〉 3D graph for delay

다양한 SNR에 따른 잡음환경에 노출된 음성신호에 대한 CE, CS, PESQ 측정결과는 <표 2>에 나타나 있다. <표 2>에 따르면 SNR 크기와 PESQ는 비례하는 것을 알 수 있다. CE의 변화는 음성구간검출을 통해 음성 구간을 얼마나 정확히 검출하고 있는지를 나타낸다.

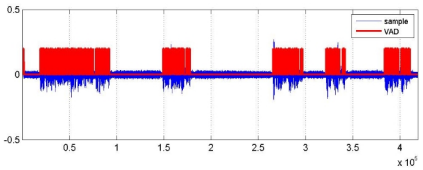
〈표 2〉 다양한 SNR에 따른 실험결과
〈Table 2〉 Experimental results due to various SNR

SNR	30dB	25dB	20dB	15dB
CE	0.881	0.879	0.873	0.868
CS	1.824	1.827	1.828	1.930
PESQ	3.735	3.379	2.988	2.632

동일한 지연인 경우에도 잡음의 변화에 따라 실제로 인지되는 음성구간은 점점 짧아지는 것을 CE의 변화를 통해 알 수 있다. SNR이 낮아지면서 자연스럽게 잡음보다 크기가 작은 음성구간은 들리지 않는 마스크효과가 발생한다. 이러한 원인이 실제로 검출되는 음성구간에 영향을 주게 되어 CE값이 감소하게 된다. <그림 6>은 잡음 환경에 따른 음성구간 검출 결과를 비교하여 나타낸다.



a) 무 잡음 신호



b) 잡음 신호(white 15dB)

〈그림 6〉 음성구간 검출 결과
〈Fig 6〉 results of voice activity detection

<그림 6>의 빨간색 네모 안에 포함된 부분이 음성으로 검출된 구간이다. 잡음 환경에서 잡음과 비슷한 음압을 갖는 음성 구간 내의 구간 들은 음성구간으로 포함될 수 있도록 하였다. 위 그림에서 나타나는 것처럼 SNR에 따라 검출되는 음성구간이 차이를 보이는 것을 알 수 있다. 이로 인해 CE, CS가 변하게 되는 것이다. <표 3>은 다음은 IP 네트워크 망변동에 의한 손실 변화에 따른 CE, CS, PESQ 측정결과를 나타낸다.

〈표 3〉 다양한 손실에 따른 실험결과
〈Table 3〉 Experimental results due to various IP network loss

손실	0%	10%	20%	30%
CE	0.878	0.870	0.869	0.875
CS	1.725	1.825	1.819	1.830
PESQ	4.095	2.794	2.694	2.222

음성구간 내의 음성손실에 따른 PESQ변화는 잡음변화와 유사하게 손실이 크면 클수록 음질이 떨어지기 때문에 PESQ도 감소하고 있다. 손실이 발생하였지만 녹음되는 음성 구간은 에너지가 존재하고 있기 때문에 음성구간에 대한 인식은 크게 변화하지 않았다.

Ⅴ. 결 론

본 논문에서는 대화형 VoIP 품질 평가 방법론 및 평가 환경을 구축하고 VoIP 엔진 기능 및 망 특성에 따라 VoIP 성능을 평가할 수 있는 검증 시스템을 제안하였다. 구축된 VoIP 검증 시스템에서는 대화형 테스트 음원 DB를 통해 다양한 잡음 환경, IP 네트워크 망에 따른 네트워크 지연 및 손실에 대한 PESQ, CE, CS를 측정하였으며, PESQ-CE-CS의 상호관계를 분석하였다.

향후 연구과제로는 초광대역(28kHz)내의 네트워크 망에서 발생하는 불연속성, 잡음, 에코 등의 문제에 따른 음질변화를 객관적으로 측정할 수 있는 구체적인 연구가 진행될 예정이다.

참 고 문 헌

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

[2] “Methods for subjective determination of transmission quality,” *ITU-T Recommendation P.800*, Aug. 1996.

[3] B. W. Wah and B. Sat, “Analyzing voice quality in popular VoIP applications,” *IEEE Multimedia archive*, vol.16, no.1, pp.46-59, Mar. 2009.

[4] B. W. Wah and B. Sat, “The design of VoIP system with high perceptual conversational quality,” *Academy publisher Journal of Multimedia*, vol. 4, no. 2, pp.49-62. Apr. 2009.

[5] K. sakhnov, E. Verteletskaya and B. Simak, “Approach for energy-based voice detector with adaptive scaling factor,” *IAENG International Journal of Computer Science*, vol. 36, no. 4, Nov. 2009.

[6] M. Liem, H.-G. Kim and O. Manck, “Algorithm of a single chip acoustic echo canceller using cascaded cross spectral estimation,” *International Workshop on Acoustic Echo and Noise Control* (*IWAENC 2003*), pp.187-189, Sep. 2003.

저자소개

조 재 만 (Cho, Jae-Man)

2011년 3월 ~ 현 재 : 광운대학교 전자융합공학과 석사과정
2011년 : 광운대학교 전파공학과 공학사

김 형 국 (Kim, Hyoung-Gook)

2007년 3월 ~ 현 재 : 광운대학교 전자융합공학과 부교수
2005년 4월 ~ 2007년 2월 : 삼성종합기술원 수석연구원
2002년 8월 ~ 2005년 3월 : 독일 베를린 공과대학교 Assistant Professor
1999년 1월 ~ 2002년 7월 : 독일 SIEMENS/Cortologic AG 책임연구원