한국ITS학회논문지
제7권, 제5호 (2008년 10월)
pp.90~96

# 텔레메틱스 단말용 음성인식을 위한 음성향상 알고리듬 및 칩 구현

# Implementation of Chip and Algorithm of a Speech Enhancement for an Automatic Speech Recognition Applied to Telematics Device

김 형 국*
(Hyoung-Gook Kim)

## 요 약

본 논문은 텔레메틱스 단말용 음성인식을 위한 음성향상 단일 칩 알고리듬을 제시한다. 제안된 방법은 잡음제거와 에코제거의 두 단계로 구성되어 있으며, 첫 단계로 크로스 스펙트럼 추정에 기반한 적응필터를 통해 에코를 제거하고, 두 번째 단계로 Generalized Gamma 분포기반의 LSA 음성추정 방식 추정을 통해 외부 배경잡음을 제거하여 음성의 음질을 향상시킨다. 적은 계산량이 요구되는 제안된 알고리즘을 토대로 구현된 단일 칩의 성능은 다양한 잡음환경에서 신호 대 잡음비율과 음성인식 평가에서 기존의 방법보다 향상된 결과를 나타내었다.

## Abstract

This paper presents an algorithm of a single chip acoustic speech enhancement for telematics device. The algorithm consists of two stages, i.e. noise reduction and echo cancellation. An adaptive filter based on cross spectral estimation is used to cancel echo. The external background noise is eliminated and the clean speech is estimated by using MMSE log-spectral magnitude estimation. To be suitable for use in consumer electronics, we also design a low cost, high speed and flexible hardware architecture. The performance of the proposed speech enhancement algorithms were measured both by the signal-to-noise ratio(SNR) and recognition accuracy of an automatic speech recognition(ASR) and yields better results compared with the conventional methods.

**Key words**: Noise reduction, echo cancellation, automatic speech recognition

† 본 연구의 하드웨어 칩 구현에 도움을 주신 독일 Ruwisch & Kollegen GmbH에 감사드립니다.
* 주저자 : 광운대학교 전파공학과 교수
† 논문접수일 : 2008년 8월 12일
† 논문수정일 : 2008년 9월 22일
† 게재확정일 : 2008년 9월 22일

# Ⅰ. 서 론

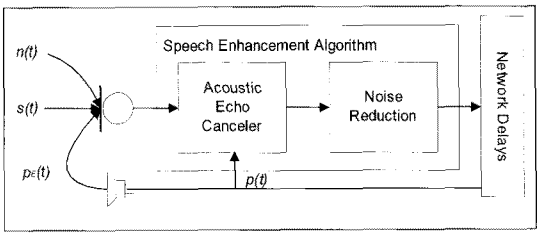
최근 컴퓨터와 무선 데이터전송의 발달에 따른 텔레메틱스 기술 응용분야가 확산됨에 따라, 손의 사용이 제한되는 운전 중인 차량환경에서 주행 중 행선지를 입력하기 위해 터치스크린을 사용하는 문자입력 방식의 불편함을 해소할 수 있는 텔레메틱스 단말용 음성인식 기술이 개발 및 사용되어 오고 있다. 주행 중 차량내부에는 엔진소리, 차량내부 공간에 의해 발생되는 수신음성신호의 에코, 그리고 차량외부에 존재하는 바람소리 등의 배경잡음 등이 음성인식 성능을 저하시킨다. 잡음환경에서의 강인한 음성인식 분야는 음성인식 연구 분야 중에 가장 주목을 받고 있으며, 인식성능을 높이기 위한 다양한 방법들이 제안 및 개발되어 오고 있다. 이러한 연구결과는 대체적으로 음성향상기법 [1], 잡음에 강인한 특징추출기법 [2] 그리고 인식모델방법 등으로 구분 지을 수 있다.

본 논문에서는 주행 중 음성인식의 성능에 순간적으로 악영향을 미치는 차량 안에서 발생하는 에코와 배경잡음을 제거하는 음성향상 알고리즘을 제안한다. 적은 계산량이 요구되는 제안된 알고리즘을 토대로 음성향상 단일 칩을 구현하여 다양한 잡음환경에서 음성인식의 성능을 측정하였다.

본 논문의 구성은 다음과 같다. Ⅱ 장에서는 제안된 음성향상 방식에 대해서 소개하며 Ⅲ 장에서는 제안된 방식의 칩 구현을 설명하고 Ⅳ 장에서는 인식실험결과를 소개하고 마지막으로 Ⅴ 장에서 결론을 맺는다.

# Ⅱ. 음성향상 알고리즘

본 논문에서 사용한 음성향상 알고리즘은 <그림 1>에 나타난 바와 같이 에코제거와 잡음제거의 두 단계로 구성되어 있으며, 첫 단계로 두 단계 크로스 스펙트럼 추정에 기반한 적응필터를 통해 에코를 제거하고, 두 번째 단계는 generalized Gamma 분포기반 [3]의 Minimum Mean-Square Error Log-Spectral Amplitude Estimator(MMSE LSA) [1] 추정을 통해 외부 주변잡음을 제거하여 음성의 음질을 향상시킨다.




<그림 1> 에코 및 잡음제거 음질향상 알고리즘 구조
<Fig. 1> Block diagram of a speech enhancement algorithm.

## 1. 에코제거

잡음신호 $n(t)$와 에코신호 $p(t)$가 음성신호 $s(t)$에 인가되면 오염된 음성신호 $y(t)=s(t)+n(t)+p(t)$를 만들어 내게 되고 퓨리에 변환을 통해 주파수 축에서 다음과 같이 표현된다.

$$Y_1(k,l)=S_1(k,l)+N_1(k,l)+P_1(k,l) \quad (1)$$

$$Y_2(k,l)=S_2(k,l)+N_2(k,l)+P_2(k,l) \quad (2)$$

여기서 $Y_1(k,l)$, $S_1(k,l)$, $N_1(k,l)$, $P_1(k,l)$은 $y(t)$, $s(t)$, $n(t)$, $p(t)$의 256 샘플에 의한 퓨리에 변환을 통한 1번째 프레임에서의 k번째 주파수 성분이 되고, $Y_2(k,l)$, $S_2(k,l)$, $N_2(k,l)$, $P_2(k,l)$은 $y(t)$, $s(t)$, $n(t)$, $p(t)$의 2048 샘플에 의한 퓨리에 변환을 통한 주파수 성분이 되며 다음과 같이 표현된다.

$$Y_1(k,l)=R_1(k,l)\exp(j\Theta_1(k,l)) \quad (3)$$

$$Y_2(k,l)=R_2(k,l)\exp(j\Theta_2(k,l)) \quad (4)$$

$$S_1(k,l)=A_1(k,l)\exp(j\Phi_1(k,l)) \quad (5)$$

$$S_2(k,l)=A_2(k,l)\exp(j\Phi_2(k,l)) \quad (6)$$

여기서 $R(k,l)$, $\Theta(k,l)$은 각각 $y(t)$의 진폭, 위상을 나타내고 $A(k,l)$, $\Phi(k,l)$은 $s(t)$의 진폭, 위상을 나타낸다.

에코경로추정을 위해 $Y_2(k,l)$와 $P_2(k,l)$의 크로스 스펙트럼과 $P_2(k,l)$와 $P_2^*(k,l)$의 크로스 스

펙트럼의 비 [5]를 아래와 같이 구한다.

$$W_2(k,l)=\frac{C_{YP}}{C_{PP}}=\frac{Y_2(k,l)P_2(k,l)}{P_2(k,l)P_2^*(k,l)} \quad (7)$$

위의 식에서 2048 샘플에 의한 퓨리에 변환 주파수는 크로스 스펙트럼 계산에서의 wrap-over 영향을 감소시키고 에코경로추정의 under-modeling에 의해 발생되는 에러를 저지한다.

식 (7)에 의해 구해진 $W_2(k,l)$은 $W_1(k,l)$로 변환되고

$$W_1(k,l)=\sum_{i=-32}^{32} W_2\left(k+\frac{i}{8},l\right)f(i) \quad (8)$$

Leakage 함수 $f(i)$에 의해 128 주파수 성분으로 전환된 $W_1(k,l)$은 (9) 식의 normalized LMS(NLMS) 알고리즘을 통해 입력된 에코신호 $p_E(k,l)$의 주파수성분 $P_E(k,l)$에 해당되는 에코경로성분 $W_1(k,l)P_1(k,l)$을 추정한다.

$$W_1(k,l+1)=W_1(k,l)+\frac{E_1(k,l)P_1(k,l)}{(1-\tau)P_1(k,l-1)+\tau P_1(k,l)P_1^*(k,l)},\quad (0<\tau<1) \quad (9)$$

추정된 에코경로성분을 입력신호의 주파수 성분으로부터 차감함으로써 일차적으로 주요 에코성분을 아래와 같이 제거할 수 있다.

$$E_1(k,l)=Y_1(k,l)-W_1(k,l)P_1(k,l) \quad (10)$$
$$=S_1(k,l)+N_1(k,l)+\{P_E(k,l)-W_1(k,l)P_1(k,l)\}$$

다음으로, 간섭제거 후단에 $E_1(k,l)$와 $P_2(k,l)$의 크로스 스펙트럼과 $P_1(k,l)$와 $P_1^*(k,l)$의 크로스 스펙트럼의 비를 $E_1(k,l)$에 곱하여 획득된 잔여 에코추정 성분을 잔여에코 성분 $E_1(k,l)$에서 차감하여 잔여 에코제거를 제거한다.

$$X(k,l)=E_1(k,l)-\frac{E_1(k,l)P_1(k,l)}{P_1(k,l)P_1^*(k,l)}E_1(k,l) \quad (11)$$

## 2. 잡음제거

잡음제거는 먼저 잡음스펙트럼추정과 MMSE LSA 음성추정을 통해 음질이 향상된 음성 성분을 획득할 수 있다. 잡음스펙트럼을 추정할 수 있는 최소잡음 스펙트럼 성분을 얻기 위해 주파수 축과 시간 축에 대해 스무딩이 적용된 전력성분을 아래와 같이 구한다.

$$X_F(k,l)=\frac{1}{2w+1}\sum_{i=-w}^{w}|X(k,l)| \quad (12)$$

$$X_T(k,l)=\alpha_T X_T(k,l-1)+(1-\alpha_T)X_F(k,l) \quad (13)$$

여기서 $X_F(k,l)$, $X_T(k,l)$은 주파수 축 스무딩으로 구해진 스펙트럼과 시간 축 스무딩으로 구해진 스펙트럼 성분이며 $\alpha_T(0<\alpha_T<1)$는 스무딩 파라미터이다.

전력스펙트럼으로부터 시간 축 프레임 C를 따라 최소잡음성분을 아래와 같이 추정한다.

$$M(k,l)=\min_{c=0,\ldots,C}\{M(k,l-c),X_T(k,l)\} \quad (14)$$

최소잡음성분과 시간 축 스무딩이 적용된 잡음전력의 비를 사용하여 음성존재 구간과 비음성존재 구간을 구별하는 VAD(voice activity detection) Z(k,l)를 아래와 같이 구한다.

$$Z(k,l)=\begin{cases}1 & \text{if } X_T(k,l)/M(k,l)\\ 0 & else\end{cases} \quad (15)$$

VAD는 잡음전력 추정갱신을 조정하며, 음성존재확률 추정값인 $p_s(k,l)$은 잡음성분 추정을 위한 최적 스무딩 함수 $\alpha(k,l)$을 획득하기 위해 아래와 같이 계산된다.

$$\text{if } Z(k,l)=1 \qquad (16)$$
$$p_s(k,l)=\alpha_p+(1-\alpha_p)p_s(k,l-1),$$
$$\alpha(k,l)=1$$

*else*
$$p_s(k,l)=(1-\alpha_p)p_s(k,l-1)$$
$$\alpha(k,l)=\alpha_a+(1-\alpha_a)p(k,l)$$

식에서 $\alpha_p(0<\alpha_p<1)$, $\alpha_a(0<\alpha_a<1)$이다.

음성이 존재한다고 가정되는 구간에서는 스무딩 함수 $\alpha(k,l)$은 1로 세트시키고 잡음추정은 즉시로 정지함으로써, 실제적인 음성구간에서 잘못된 잡음 추정을 방지할 수 있다. 음성이 존재하지 않는다고 가정되는 구간에서는 음성존재확률 추정값은 높은 스무딩 파라미터를 이용하여 회귀적으로 감소되어 잡음이 존재하는 구간의 시작부분에서 천천히 잡음 추정을 수행한다. 스무딩 함수를 이용하여 최종적으로 아래와 같이 잡음전력이 추정된다.

$$\lambda_N(k,l)=\alpha(k,l)\lambda_N(k,l-1) + (1-\alpha(k,l))|X(k,l)| \qquad (17)$$

MMSE LSA에 기반한 잡음제거 이득 $G(k,l)$은 a priori SNR $\xi(k,l)$과 a posteriori SNR $\gamma(k,l)$을 사용하여 아래와 같이 표현된다.

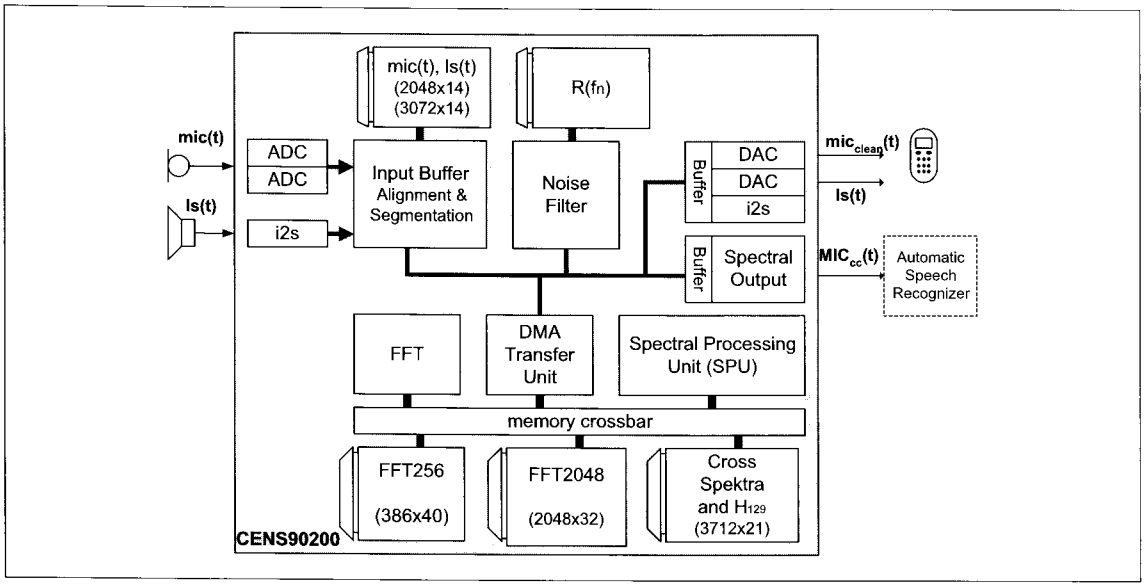
$$G(k,l)=\frac{\xi(k,l)}{1+\xi(k,l)}\exp\left(0.5\int_{t=v(k,l)}^{\infty}\frac{e^{-t}}{t}dt\right) \qquad (18)$$

$$\xi(k,l)=\max\left[\beta\frac{\widetilde{A}(k,l-1)}{\lambda_N(k,l)}+(1-\beta)(\gamma(k,l)-1),\xi_{\min}\right] \qquad (19)$$

$$v(k,l)=\frac{\xi(k,l)}{1+\xi(k,l)}\gamma(k,l) \qquad (20)$$

$$\gamma(k,l)=\begin{cases}\dfrac{|X(k,l)|}{\lambda_N(k,l)}, & \text{if } |X(k,l)|>\lambda_N(k,l)|\\ 1 & else\end{cases} \qquad (21)$$

본 논문에서는 최근에 음성추정을 위해 효과적으로 사용되고 있는 다음과 같은 generalized Gamma 분포기반 [3]의 LSA 음성추정 방식을 사용한다.




<그림 2> 에코 및 잡음제거 음질향상 칩 아키텍처

<Fig. 2> Architecture of a speech enhancement single chip for echo cancellation and noise reduction.

$$\widetilde{A}(k,l)=\exp\left[\frac{\int_0^{\infty}\log(a)p_A(a)e^{-\frac{a^2}{\lambda_N}}I_0\left(\frac{2|X(k,l)|}{\lambda_N}a\right)da}{\int_0^{\infty}p_A(a)e^{-\frac{a^2}{\lambda_N}}I_0\left(\frac{2|X(k,l)|}{\lambda_N}a\right)da}\right] \quad (22)$$

$$p_A(a)=\frac{\kappa a^{\nu}}{\Gamma(\nu)}a^{\kappa\nu}\exp(-\alpha a^{\kappa})$$

음성향상 스펙트럼은 오염된 음성신호와 잡음제거이득의 곱인 $|\widetilde{S}(k,l)|=G(k,l)\cdot|X(k,l)|$를 통해 최종적으로 구할 수 있다.

## Ⅲ. 음성향상 칩 디자인

음향에코와 배경잡음을 제거하는 음질향상 구현 칩인 Speech Enhancement Chip(SEC)은 <그림 2>에 나타나 있으며 독립실행을 위해 필요한 모든 구성요소를 포함한다.

마이크에 입력되는 오염된 입력신호 $y(t)$=mic(t)와 스피커신호 $p(t)$=ls(t)는 두 개로 결합된 14 비트 G.711 규격 음성코덱에 의해 샘플링되거나 12S 인터페이스에 의해 디지털방식으로 공급된다. 입력버퍼에서는 정렬된 mic(t)와 ls(t)의 연속되는 샘플스트림을 퓨리에변환(FFT)을 위해 256개의 샘플과 2048개의 샘플로 각각 분할하고 스피커신호를 마이크 입력신호에 대하여 1024개의 신호샘플로 지연시킨다. 분할된 음성 세그멘트는 FFT, SPU와 잡음제거 신호처리 유닛에 의해 처리되어지며 신호처리 자체는 할애된 하드웨어 유닛과 스펙트럼 신호처리 유닛(SPU)사이에서 분산 처리된다. 퓨리에 변환과 잡음처리 필터는 하드웨어에 직접적으로 매핑되어 신호처리 시간이 바로 마이크신호의 지연에 영향을 미치고 알고리즘의 남아있는 부분의 신호처리는 SPU에서 수행된다. SPU는 스펙트럼 신호처리를 위해 최적화된 RISC 스타일 적용 특정 프로세서로서 산술연산, 분할, 메모리 트렌스퍼, 제로 사이클 컨디셔널 점프 등을 병렬로 처리한다.

잡음제거 유닛은 에코제거 후 획득된 128 주파수 진폭성분에서 최대 12298 클럭 사이클을 사용하여 잡음제거 신호처리를 수행한다. 에코와 잡음이 최종적으로 제거된 음질향상 진폭은 IFFT에 의해 아날로그 신호로 전환되어 지거나 음성특징 추출에 적용되어 음성인식 시스템에 사용된다.

<표 1> 음질향상 칩의 technical characteristics
<Table 1> Technical characteristics of integrated speech enhancement chips.

| 알고리즘 | SEC | CS [4] | MSM [5] | PSB [6] |
|---|---|---|---|---|
| fclk[Mhz] | 12.29/16.93 | 20.48 | 19.2 | 17/34.56 |
| I[mA@V] | 34/43@3.3 | 60@5 | 35@3 | 30/50@3.3 |
| fs[kHz] | 8/11.025 | 8 | 8 | 8 |
| max AEC | 128+128 | 63.5 | 59 | 50-129/96 |
| length[ms] | 92.8+92.8 | | | |

위에서 설명된 아키텍처는 에코 및 잡음을 제거하는 음질향상 알고리즘을 매우 능률적으로 실행시킨다. 즉, 모든 입력신호에 대해 최대 1450 클럭사이클을 필요로 하며, 이것은 8kHz 샘플링레이트에서 11.6 MHz의 최소 수행주파수에 해당된다.

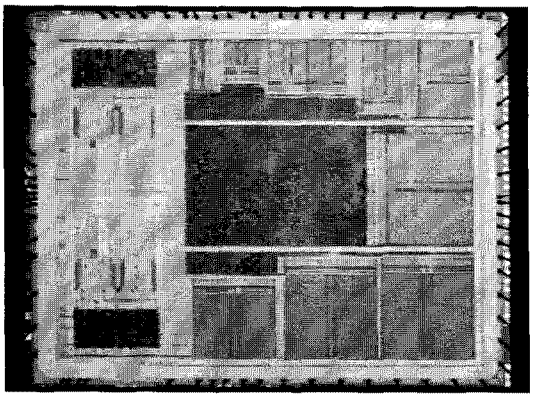
게다가, 제안된 알고리즘의 아키텍처는 에코 및 잡음제거 알고리즘을 Intellectual Property(IP) core와 같은 다른 디자인으로 쉽게 통합되어 질 수 있도록 최대 portability에 최적화되어 있다. 그러므로 모든 메모리 액세스는 유용 메모리 throughout를 효과적으로 이등분 시키는 2개의 클럭 사이클로 주어지게 되나 통합메모리의 어떤 타입도 사용될 수 있도록 안전하게 설계되어 있다.

제안된 아키텍처를 전력소비와 수행주파수에 대해서 다른 기존의 칩과 비교하면 <표 1>과 같다.

SEC의 dies size는 0.35μm 3 layer metal CMOS 프로세스의 36 $mn^2$이고 primitive logic gate count는 81,516 nand2 gate equivalents로 <그림 3>과 같이 구성된다.

## Ⅳ. 실험 및 결과고찰

본 논문에서는 제안된 음질향상 알고리즘의 음질평가를 위해 에코 및 잡음 신호처리 전과 후의 신호에 대한 SNR 개선과 음성인식 성능을 측정하였다.



<그림 3> SEC의 Die photo
<Fig. 3> Die photo of the SEC

<표 2> 기존방식과 음질향상 칩의 SNR 개선 결과 비교
<Table 2> Comparison of SNR improvement of different speech enhancement algorithms

| | Echo(15 dB) White Noise (6dB) | Echo(15 dB) Street Noise (6dB) | Echo(15 dB) Color Noise (6dB) |
|---|---|---|---|
| SS+CL | 8.26 | 7.85 | 7.09 |
| LSA+MSM | 8.83 | 8.39 | 7.31 |
| SEC | 9.35 | 8.53 | 7.87 |

<표 2>는 에코신호와 3가지 타입(백색잡음, 거리잡음, 컬러잡음)의 배경잡음을 스튜디오 음성신호에 인위적으로 혼합하여 음성향상 칩을 통해 신호처리 된 음성신호에 대한 SNR 개선 측정결과이다.

<표 2>의 결과를 고찰해 보면 제안된 방식의 음성향상 알고리즘이 다양한 잡음과 에코가 섞인 음성신호로부터 기존의 방식에 비해 보다 향상된 SNR 개선을 보여줌을 알 수 있다.

<표 3>은 음성신호의 분할된 각 프레임에서 잡음이 제거된 음성 진폭으로부터 각 총 39개의 음성 특징값(13 Mel-frequency cepstral coefficients, 13 deltas, 13 accelerations)을 추출하여 Markov 모델의 확률적 추정에 의한 기법을 도입한 HMM(hidden Markov model) 음성인식 시스템에 적용한 음성인식 결과를 나타내었다.

<표 3> 기존의 방식과 비교된 SEC 음성인식 결과
<Table 3> Comparison of speech recognition result of SEC with different speech enhancement algorithms

| | Set A | Set B | Set C |
|---|---|---|---|
| SS | 88.51 | 87.57 | 85.56 |
| LSA | 90.90 | 89.60 | 88.35 |
| SEC | 91.26 | 90.10 | 89.72 |

제안된 SEC 방식은 기존의 SS나 LSA방식에 비해서 Aurora 2의 모든 데이터베이스에서 향상된 인식성능을 보임을 알 수 있었다.

## V. 결 론

본 논문에서는 주행 중 음성인식성능을 향상시키기 위해 에코와 배경잡음을 제거하는 음성향상 알고리즘을 제안하였다. 적은 계산량이 요구되는 제안된 알고리즘을 토대로 구현된 단일 칩의 성능을 다양한 잡음환경에서 신호 대 잡음비율과 음성인식 성능을 평가한 결과 제안된 알고리즘이 기존의 방법 [1], [6-8] 보다 향상된 결과를 나타내었다. 제안된 방식의 강인성은 실제 차량 주행환경에서 매우 유용하게 사용될 수 있으리라 생각된다.

## 참고문헌

[1] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator", IEEE Trans. Acoustics, Speech and Signal Proc. Vol. 33, No.2, pp. 443-445, Dec. 1985.

[2] U. H. Yapanel and J. h. L. Hansen, "New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition", Proc. Eurospeech 2003, Sep. 2003.

[3] M. J. F. Gales, "Model Based Techniques for Robust-Speech Recognition", Ph. D. Dissertation, University of Cambridge, 1995.

[4] R. C. Hendriks, J. S. Erkelens, J. Jensen, and R.

Heusdens, "Minimum Mean-Square Error Amplitude Estimators for Speech Enhancement under the Generalized Gamma Distribution," Proc. IWAENC, Sep. 2006.

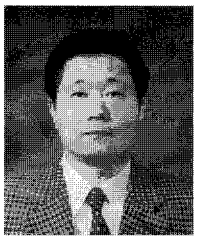
[5] Dietmar Ruwisch: "Verfahren und Vorrichtung zur Elimination von Lautsprechersignalen aus Mikrophonsignalen", Patent De 100 43 064 A1, 2000.

[6] Cirrus Logic: "CS6422 Enhanced Full-Duplex Speakerphone IC" (Datasheet), July 2001.

[7] OKI Semiconductor: "MSM7731-02 Voice Signal Processor" (Datasheet), May 2001

[8] Infinion Technologies: "Acoustic Echo Canceller ACE PSB2170 Version 2.1" (Datasheet), 1999.

## 저자소개



**김 형 국 (Kim, Hyoung-Gook)**

2007년 3월~현재 : 광운대학교 전파공학과 조교수
2005년 4월~2007년 2월 : 삼성종합기술원 수석연구원
2002년 8월~2005년 3월 : 독일 베를린 공과대학교 Adjunct Professor
1999년 1월~2002년 7월 : 독일 Cortologic AG 책임연구원