한국음향학회지 제37권 제6호 pp. 512~517 (2018)
The Journal of the Acoustical Society of Korea Vol.37, No.6 (2018)
https://doi.org/10.7776/ASK.2018.37.6.512

pISSN : 1225-4428
eISSN : 2287-3775

배경잡음 및 패킷손실에 강인한 voice-over-IP 수신단 기반 음질향상 기법

Robust speech quality enhancement method against background noise and packet loss at voice-over-IP receiver

김지연,[1] 김형국[1†]

(Gee Yeun Kim[1] and Hyoung-Gook Kim[1†])

[1]광운대학교 전자융합공학과

(Received September 18, 2018; revised November 5, 2018; accepted November 22, 2018)

초 록: 음성 품질의 향상은 통신 분야의 주요 관심사이다. 본 논문에서는 VoIP(Voice-over-IP) 수신부에서의 배경잡음 및 패킷손실에 강인한 음질향상 방식을 제안한다. 제안된 방식에서는 하이브리드 마르코프 체인 기반 네트워크 지터추정, 추정된 지터를 이용한 적응적 플레이아웃 스케줄링, 그리고 진폭 및 위상 복원 기반의 음성 향상 방식 등을 결합하여 IP 네트워크를 통해 VoIP 수신부에 도착하는 음성신호의 품질을 향상시킨다. 실험결과는 제안된 방식이 송신부의 인코딩 전에 음성신호에 추가된 잡음을 제거하고 불안정한 네트워크 환경에서 양질의 음성을 제공하는 것을 확인할 수 있다.
핵심용어: 음성인터넷 프로토콜, 지터추정, 적응적 플레이아웃 스케줄링, 패킷손실 은닉, 음성 향상

ABSTRACT: Improving voice quality is a major concern in telecommunications. In this paper, we propose a robust speech quality enhancement against background noise and packet loss at VoIP (Voice-over-IP) receiver. The proposed method combines network jitter estimation based on hybrid Markov chain, adaptive playout scheduling using the estimated jitter, and speech enhancement based on restoration of amplitude and phase to enhance the quality of the speech signal arriving at the VoIP receiver over IP network. The experimental results show that the proposed method removes the background noise added to the speech signal before encoding at the sender side and provides the enhanced speech quality in an unstable network environment.
Keywords: VoIP (Voice-over-IP), Jitter estimation, Adaptive playout scheduling, Packet loss concealment, Speech enhancement
PACS numbers: 43.72.Dv, 43.72.Kb

I. 서 론

최근 들어 인터넷 통신망의 고도화와 스마트 기기의 보급으로 VoIP(Voice-over-IP)를 이용한 영상 및 음성통화 서비스가 제공되고 있다. 그러나 VoIP 기반의 음성통화에서는 기존의 전화와 달리 IP 네트워크 환경의 변동으로 인해 발생하는 지연, 지터 및 패킷 손실과 같은 전송 장애가 음질에 영향을 주고, 인코딩 전에 음성 신호에 추가된 잡음은 인코딩 유형에 따라 예측할 수 없는 음성품질 결과를 초래한다.

이러한 문제들을 해결하기 위해 수신부에 불규칙한 간격으로 도착한 패킷들로 인해 발생하는 버퍼링 지연을 감소시키고 패킷손실을 은닉하기 위한 다양한 방식들[1]이 개발되어 오고 있다. 또한, 딥러닝 방식을 이용하여 배경잡음을 제거하고 통화의 음질을 향상시키기 위한 연구[2,3]가 독립적으로 진행되고 있다. 그러나 아직 패킷손실은닉과 잡음제거를 결합하여 보다 더 나은 고품질의 음성을 획득하기 위한 연

†**Corresponding author:** Hyoung-Gook Kim (hkim@kw.ac.kr)
Department of Electronics Convergence Engineering, Kwangwoon University, 20 Gwangun-ro, Nowon-gu, Seoul 01897, Republic of Korea
(Tel: 82-2-940-5574, Fax: 82-2-913-5006)
"이 논문은 2018년도 한국음향학회 음성통신 및 신호처리 학술대회에서 발표하였던 논문임."

구는 아직 미비하다.

본 논문에서는 패킷 손실 은닉을 통해 수신된 음성신호에 여전히 존재하는 잔여 잡음을 제거하고 보다 더 향상된 고품질의 음성을 획득하기 위해 저연산량으로 패킷손실은닉과 딥러닝 방식을 이용한 잔여잡음 제거를 결합하는 방식을 제안한다.

본 논문의 구성은 다음과 같이 구성되어 있다. II장에서는 제안된 방식에 대해서 설명하고, III장에서는 제안된 방식의 성능을 확인하기 위한 실험결과를 제시한다. 그리고 IV장에서는 결론을 서술한다.

II. VoIP 수신단 기반 음질향상

Fig. 1은 본 논문에서 제안하는 VoIP 수신단 기반 음질향상을 나타내는 블록도이다.

송신부에서는 마이크로 입력되는 송신자의 음성을 인코딩과 패킷화를 통해 수신부에 전송한다.

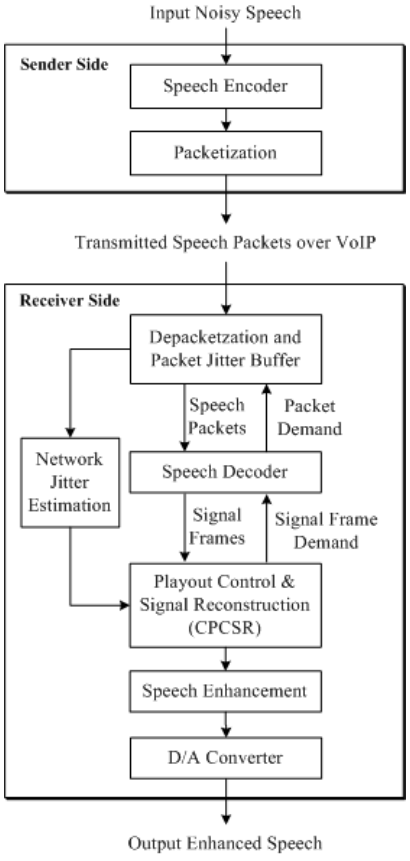

Fig. 1. Overall flow chart of the proposed VoIP receiver-based speech quality enhancement.

송신부와 수신부사이의 IP 네트워크에서는 불규칙한 네트워크 지터로 인해 버퍼링 지연과 패킷손실이 발생된다. 음성패킷이 수신부에 도착하면, 지터 버퍼에 음성패킷을 저장하고 지터 버퍼의 패킷들을 시퀀스번호에 맞게 정렬한다. 지터 추정부에서는 수신된 패킷의 해더정보를 이용하여 IP 네트워크 지터를 추정하고, 지터 버퍼에 저장된 음성패킷은 디코더를 통해 음성패킷을 요청받아 신호프레임으로 전환하여 적응적 플레이아웃 조정 및 신호 재구성부(Adaptive Playout Control and Signal Reconstruction, A-PCSR)로 전달된다. A-PCSR에서는 패킷손실 및 병합과정을 통해 신호를 재구성하고, 추정된 지터를 통한 적응적인 플레이아웃 조정을 거쳐 버퍼링 지연과 패킷손실을 최소화 시켜 음성프레임을 음질 향상부에 전달한다. 음질 향상부에서는 음성신호의 진폭과 위상 모두를 복원시킴으로써 송신자의 음성신호에 추가된 잔여잡음을 제거하여 통화음질을 향상시키고 이를 D/A (Digital-to-Analog) 변환기를 통해 출력한다.

2.1 IP 네트워크 지터추정

Fig. 2는 지터추정 과정을 나타낸 블록도이다.

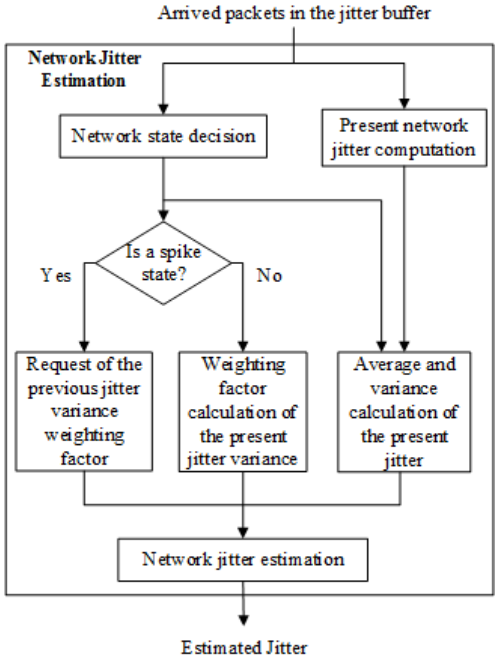

Fig. 2. Block diagram of network jitter estimation.

제안된 방식에서는 현재 수신단에 도착한 패킷의 도착 및 생성시간을 비교하여 패킷이 통과한 현재 네트워크 지터를 계산하고, 하이브리드 마르코프 체인 방식[4]을 통해 지터의 변화범위가 크지 않은 안정적인 정상상태와 지터의 변화가 큰 불안정한 스파이크 상태로 구분하여 현재 IP 네트워크의 환경을 추정한다. Eq. (1)을 이용하여 정상상태에서 스파이크 상태로 변화 시에 지터 $j_{i,k}$의 평균 $m_{i,k}$과 분산 $c_{i,k}$ 그리고 분산가중치 값 $\beta_{i,k}$을 저장하고 스파이크 상태에서는 이를 갱신하지 않으며, 스파이크에서 정상상태로 복원 시 저장된 지터의 평균과 분산 그리고 분산가중치 값을 바로 사용하여 스파이크에서 발생하는 지터추정 오류를 최소화한다.

$$j_{i,k} = m_{i,k} + \beta_{i,k} \cdot c_{i,k} . \quad (1)$$

2.2 적응적 플레이아웃 조정 및 신호 재구성

Fig. 3은 A-PCSR의 블록도를 나타낸다.

먼저, 지터 버퍼에서 디코딩 된 각 신호 프레임은 200 ms 길이의 신호 프레임 버퍼에 저장된다.

A-PCSR에서는 손실된 패킷을 은닉하고, 손실은닉 이후 정상 수신된 음성프레임 출력 시에 이전 손실은닉으로 생성된 음성프레임과의 불연속점을 제거하는 병합과정을 수행한다. 손실은닉 및 병합이 아닌 경우에는 추정된 지터를 적용한 적응적인 플레이아웃 스케줄링을 통해 버퍼링 지연과 패킷손실을 최소화한다. 즉, 추정된 지터와 수신부에 존재하는 음성프레임의 길이를 비교하여 네트워크 상황에 맞게 압축 및 정상출력을 판단함으로써 버퍼링 지연을 줄이고, 적응적인 피치 comb 필터[5]기반의 음성분류 결과를 이용해 묵음 및 잡음구간에서만 신호의 압축을 수행함으로써 음질을 향상시킨다. 위의 판단 모드를 통해 처리된 샘플 신호는 재생을 위해 20 ms 길

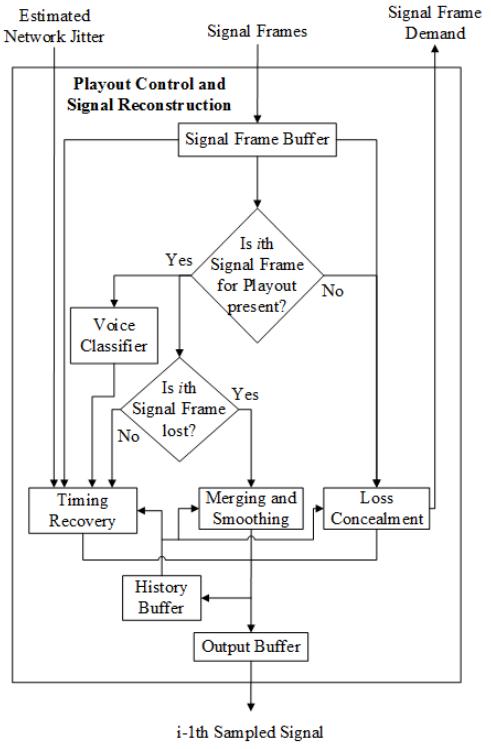

Fig. 3. Block diagram of adaptive playout control and signal reconstruction.

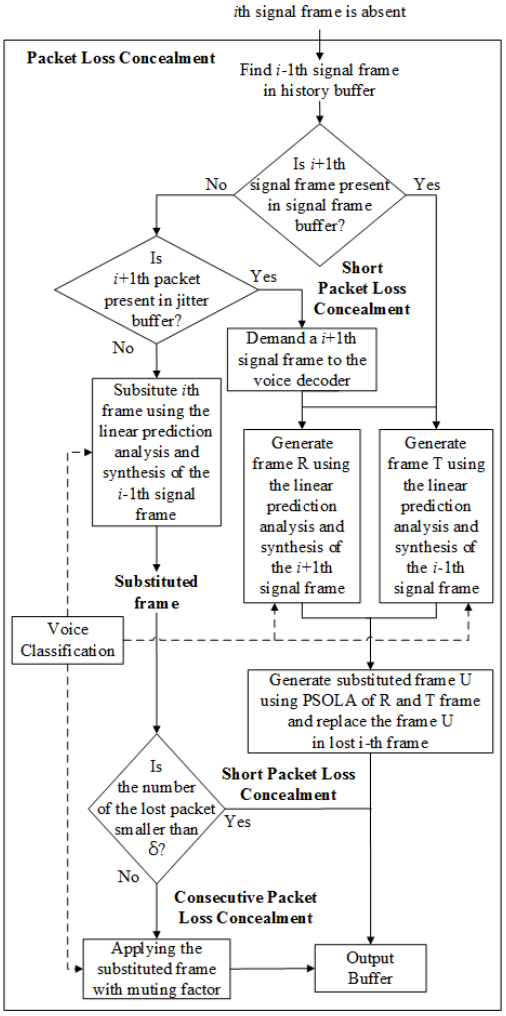

Fig. 4. Overall flow chart of packet loss concealment.

이의 출력 버퍼와 히스토리 버퍼에 입력 및 저장된다.

Fig. 4는 패킷 지연과 복잡성을 낮춰 높은 품질과 자연스러운 음질을 제공하는 광대역 음성 코덱인 G.722에 회귀적인 선형예측 분석과 합성(LPAS: Linear Prediction Analysis and Synthesis)을 기반으로 추정된 피치 주기를 단구간 및 장구간 패킷은닉에 적용함으로써 음질을 보다 더 향상시킬 수 있는 패킷손실 은닉 알고리즘의 블록도를 나타낸다. 단구간 패킷은닉에서는 추정된 피치 주기를 통해 손실 패킷들의 복원을 위한 스무딩한 여기신호를 생성하고, 단구간 패킷 손실 구간 복원 시 작은 음성 구간을 반복하였을 때 생성될 수 있는 부자연스러운 인공음을 효과적으로 줄인다. 만일 연속적인 패킷들이 손실된다면, 장구간 패킷은닉을 통해 이전 합성된 신호를 선형예측필터에 회귀적으로 입력하여 새로운 여기 신호를 생성한다. 이렇게 생성된 여기 신호를 필터링하여 재구성된 신호로 합성하고 복원된 구간동안 점차적으로 소리를 줄임으로써 사용자 청각에 불편함을 주는 금속음을 자연스럽게 제거한다.

2.3 진폭 및 위상 복원 기반의 음성 향상

Fig. 5는 패킷손실이 은닉되었지만 잔여잡음을 포함하고 있는 음성신호로부터 음질을 향상시키는 방식을 나타낸다. 즉, 저연산량의 신호처리 기법을 통해 음성 스펙트럼의 진폭과 위상 성분으로부터 잔여잡음을 제거하는 방식을 적용하였다.

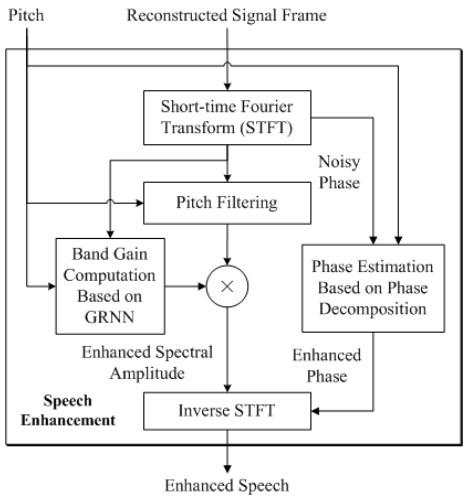

Fig. 5. Block diagram of the proposed speech enhancement.

먼저, 잡음이 제거된 음성 진폭 성분을 획득하기 위해서는 잡음환경에 노출된 음성신호로부터 STFT (Short-Time Fourier Transform)을 통해 획득된 진폭성분에 피치 필터링을 적용하여 고주파 하모닉파들 사이에 존재하는 미세한 잡음을 제거한 후 진폭추정 이득값을 적용한다.[6] 진폭추정 이득값 g_f은 스펙트럼을 22개의 Bark 스케일 대역밴드로 분할하고, 분할된 대역밴드로부터 추출한 42개의 특징값들을 보다 단순한 구조로 장단기기억 신경망과 같이 연속된 데이터의 관계를 학습할 수 있다는 이점을 갖는 두 개의 게이트형 순환 신경망(Gated Recurrent Neural Networks, GRNN)에 적용함으로써 다음과 같이 계산한다.

$$\widetilde{h_l^f} = ReLU(W \cdot [r_f^l \cdot h_{l-1}^f, x_f^l]), \tag{2}$$

$$h_f^l = (1 - z_f^l) \cdot h_f^{l-1} + z_l^f \cdot \tilde{h}_f^l, \tag{3}$$

$$g_f = \sigma(W_{h^l} \cdot h_f^l), \tag{4}$$

여기서 l은 레이어 인덱스, f는 프레임 인덱스, W는 모델의 가중치, $\sigma(\cdot)$는 sigmoid 함수를 나타낸다. 그리고 $x_f^l, z_f^l, \tilde{h}_f^l, r_f^l$은 각각 특징값, 업데이트 상태, 네트워크 상태, 리셋 상태를 의미한다.

그리고 피치 필터링의 계수는 대역별 피치 상관값과 대역별 이득값을 비교함으로 최적값을 적용한다.

음성신호가 잡음 환경에 노출될 때 위상 또한 잡음에 의해 오염되기 때문에 본 논문에서는 진폭성분에서의 잡음제거뿐만 아니라, 추가적으로 위상분해 방식[7]을 적용하여 위상에서도 잡음을 제거함으로써 음질을 개선하였다.

Fig. 6은 이러한 위상 추정의 블록도를 나타낸다.

본 논문에서는 먼저 ZZT(Zeros of the Z-Transform)를 통해 잡음에 오염된 유성음 프레임의 각 하모닉 위상을 포락선 위상과 음성 위상으로 분해한다. 그리고 ZZT의 전극 모델 파라미터를 사용하여 포락선 위상을 추정한 후에 오염된 입력 위상에서 추정한 포락선 위상을 차감하여 음성 위상을 구한다.

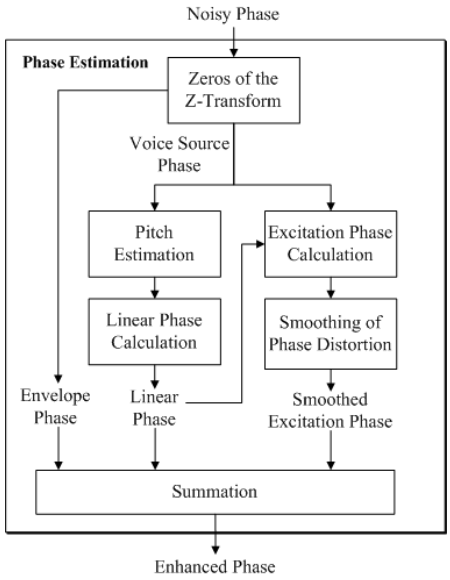

Fig. 6. Block diagram of the phase estimation.

음성 위상을 구성하는 선형위상은 인접 프레임의 선형 위상과 프레임의 기본 주파수에 의존적이며 음성신호의 고유한 위상을 포함하는 반면 여기 위상은 잡음환경에 민감하게 반응하며 변화가 많은 성분이다. 이러한 여기 위상을 잡음에 오염된 음성 위상에서 분리하고 추정된 선형이상과 포락선 위상을 모두 합하여 음성 신호의 하모닉한 구조가 복원된 향상된 위상을 획득할 수 있다.

복원된 진폭과 위상 성분에 대해 ISTFT(Inverse Short-Time Fourier Transform)를 적용하여 시간축의 향상된 음성 신호가 최종적으로 출력된다.

III. 실험 및 결과

제안한 배경잡음 및 패킷손실에 강인한 음질 향상 방법의 성능을 측정하기 위하여 session initiation protocol signaling, audio data transport, network traffic emulator 등의 모듈로 구성된 VoIP 테스트베드를 구축하고 network traffic emulator를 통해 여러 가지 네트워크 환경을 생성하였다. 이 네트워크를 통해 수신부로 음성패킷이 전송되고, 전송된 음성패킷은 VoIP 수신부의 음질 향상 시스템을 통해 음성파일로 출력된다. 음성파일은 피실험자 10명(남성 5명, 여성 5명)을 대상으로 MOS(Mean Opinion Score) 방식 기반의 청취 평가 실험에 사용되었다. 실험에는 4명의 남성과 4명의 여성 화자가 발성하는 10시간 길이의 음성신호를 16 KHz로 샘플링된 음성샘플을 사용하였으며, 음성샘플에는 배경잡음이 존재하는 환경에서의 대화내용을 포함하였다. 실험에 사용된 거리잡음은 segmental SNR(Signal to Noise Ratio) 0 dB부터 15 dB까지 5 dB의 간격으로 혼합하여 적용하였다. 실험을 위해 구성된 4개의 네트워크 환경에 대한 정보는 Table 1과 같다. Table 1에서 AND, VND, MJ, NPL은 각각 average of network delay, variance of network delay, maximum jitter, network packet loss를 나타낸다.

Table 1. Statics of network trace.

Test group	AND (ms)	VND (ms)	MJ (ms)	NPL (%)
A	35.53	11.56	145	0
B	24.82	8.30	48	1.95
C	78.83	30.13	359	2.13
D	78.52	31.27	375	4.03

Table 2. Averaged MOS scores.

Test group	Method	TBD (ms)	PLR (ms)	MOS
A	CM	52.38	0.53	3.42
	PM	51.26	0.45	3.54
B	CM	28.31	0.07	3.32
	PM	27.35	0.06	3.44
C	CM	60.62	2.52	3.08
	PM	57.79	2.35	3.28
D	CM	54.12	2.76	2.93
	PM	51.94	2.58	3.14
Averaged total	CM	48.86	1.47	3.19
	PM	47.09	1.36	3.35

성능비교를 위해 기존방식(Conventional Method, CM)과 제안방식(Proposed Method, PM)의 평균 버퍼링 지연(Total Buffering Delay, TBD), 평균 패킷 손실률(Packet Loss Rate, PLR)의 측정 값, MOS방식을 사용한 청취 평가 실험결과를 Table 2에 제시하였다. 기존방식과 제안한 방식의 차이점은 다음과 같다: 1) 지터 추정을 위해 제안방식에서는 하이브리드 마르코프 체인 방식을 통해 네트워크를 정상상태와 스파이크 상태로 구분하는 반면에 기존방식에서는 지터 변화를 정의된 스파이크 문턱값을 사용하여 구분하였다; 2) 음성 및 비음성 신호 분류를 위해 제안방식

에서는 적응적인 피치 comb 필터 방식을, 그리고 기존방식에서는 단구간 에너지, 자기상관도와 영교차점 비율의 특징값을 결합하여 적용하였다; 3) 플레이아웃 조정을 위해 제안방식에서는 추정된 지터와 패킷 손실 은닉 방식을 적용하여 신호의 압축과 정상출력을 판단하는 반면에 기존방식에서는 압축 및 신장, 그리고 정상출력 등의 세 가지로 구분하였다; 4) 음질 향상을 위해 제안방식에서는 GRNN을 게인 검출에 적용하였고, 기존방식에서는 최소 평균제곱 오차 추정을 위해 로그 스펙트럴 진폭 추정 방식을 사용하였다.

실험결과를 통해 본 논문에서 제안한 방식이 실험군 A, B, C, D 모두에서 비교된 방식보다 높은 MOS 값으로 우수한 통화품질을 제공함을 알 수 있다. 그리고 추가적으로 실험군 4개를 대상으로 잔여잡음 제거방식의 전과 후에 대한 평균적인 MOS 값을 측정하였다. 실험결과를 통해 기존방식에서는 잔여잡음 제거 방식 전에서는 2.98를 획득하였고, 제안방식에서는 3.03을 획득함을 알 수 있었다. 즉, 제안방식에서는 잔여잡음을 제거하는 방식을 통해서 MOS 값이 0.32, 기존방식에서는 0.21이 증가되었다.

IV. 결 론

본 논문에서는 VoIP 수신부에서 배경잡음 및 패킷손실에 강인한 음질향상 방법을 제안하였다. 청취실험을 통해 획득된 MOS값을 통해 적응적 플레이아웃 조정 및 신호 재구성과 잔여잡음 제거를 결합함으로써 보다 향상된 통화 품질이 제공될 수 있음을 확인할 수 있었다.

향후에는 지터추정 및 패킷손실 은닉과 잔여제거잡음 모두에 컨벌루션 네트워크를 적용함으로써 보다 더 저연산량으로 고품질의 음성을 달성하는 방식에 관해 심층적으로 연구하고자 한다.

감사의 글

본 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2018R1D1A1B07041783).

References

1. B. H. Kim, H.-G. Kim, J. Jeong, and J. Y. Kim, “VoIP receiver-based adaptive playout scheduling and packet loss concealment technique,” IEEE Trans. on Consumer Electronics, **59**, 250-258 (2013).
2. Y, Xu, J. Du, L.-I. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal Processing Letters, **21**, 65-68 (2014).
3. A. Kumar and D. Florencio, “Speech enhancement in multiple noise conditions using deep neural networks,” Proc. Interspeech, 738-3742 (2016).
4. Z. Zhao, L. Guardalben, M. Karimzadeh, J. Silva, T. Braun, and S. Sargeno, “Mobility prediction-assisted over-the-top edge prefetching for hierarchical VANETs,” IEEE J. Selected Areas in Communication, 1786-1807 (2018).
5. W. Jin, X. Liu, M. S. Scordilis, and L. Han, “Speech enhancement using harmonic emphasis and adaptive comb filtering,” IEEE Trans. Audio, Speech, and Language Processing, **18**, 356-368 (2010).
6. J.-M. Valin, “A hybrid DSP/deep learning approach to real-time full-band speech enhancement,” arxiv: 1709.08243v3 (2017).
7. H.-G. Kim and J. Y. Kim, “Adaptive single-channel speech enhancement method for a Push-To-Talk enabled wireless communication device,” IEICE Trans. on Communications, **E99-B**, 1745-1753 (2016).

저자 약력

▸ 김 지 연 (Gee Yeun Kim)

2018년 2월: 광운대학교 전자융합공학과 학사
2018년 3월 ~ 현재: 광운대학교 전자융합공학과 석사과정

▸ 김 형 국 (Hyoung-Gook Kim)

1999년 ~ 2002년: 독일 SIEMENS/ Cortologic AG 책임연구원
2002년 ~ 2005년: 독일 베를린 공과대학교 Assistant Professor
2005년 ~ 2007년: 삼성종합기술원 수석연구원
2007년 3월 ~ 현재: 광운대학교 전자융합공학과 교수