한국음향학회지 제40권 제2호 pp. 161~168 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.2 (2021)
https://doi.org/10.7776/ASK.2021.40.2.161

pISSN : 1225-4428
eISSN : 2287-3775

생성적 적대 신경망을 이용한 음향 도플러 기반 무 음성 대화기술

An acoustic Doppler-based silent speech interface technology using generative adversarial networks

이기승[1†]

(Ki-Seung Lee[1†])

[1]건국대학교 전기전자공학부

(Received January 18, 2021; accepted March 2, 2021)

초 록: 본 논문에서는 발성하고 있는 입 주변에 40 kHz의 주파수를 갖는 초음파 신호를 방사하고 되돌아오는 신호의 도플러 변이를 검출하여 발성음을 합성하는 무 음성 대화기술을 제안하였다. 무음성 대화 기술에서는 비 음성 신호로부터 추출된 특징변수와 해당 음성 신호의 파라메터 간 대응 규칙을 생성하고 이를 이용하여 음성신호를 합성하게 된다. 기존의 무 음성 대화기술에서는 추정된 음성 파라메터와 실제 음성 파라메터간의 오차가 최소화되도록 대응규칙을 생성한다. 본 연구에서는 추정 음성 파라메터가 실제 음성 파라메터의 분포와 유사하도록 생성적 적대 신경망을 도입하여 대응 규칙을 생성하도록 하였다. 60개 한국어 음성을 대상으로 한 실험에서 제안된 기법은 객관적, 주관적 지표상 으로 기존의 신경망 기반 기법보다 우수한 성능을 나타내었다.

핵심용어: 무음성 대화기술(Silent Speech Interface, SSI), 생성적 적대 신경망(Generative Adversarial Networks, GAN), 초음파 도플러, 음성합성(Speech synthesis)

ABSTRACT: In this paper, a Silent Speech Interface (SSI) technology was proposed in which Doppler frequency shifts of the reflected signal were used to synthesize the speech signals when 40kHz ultrasonic signal was incident to speaker's mouth region. In SSI, the mapping rules from the features derived from non-speech signals to those from audible speech signals was constructed, the speech signals are synthesized from non-speech signals using the constructed mapping rules. The mapping rules were built by minimizing the overall errors between the estimated and true speech parameters in the conventional SSI methods. In the present study, the mapping rules were constructed so that the distribution of the estimated parameters is similar to that of the true parameters by using Generative Adversarial Networks (GAN). The experimental result using 60 Korean words showed that, both objectively and subjectively, the performance of the proposed method was superior to that of the conventional neural networks-based methods.

Keywords: Silent speech interface, Generative adversarial networks, Ultrasonic Doppler, Speech synthesis

PACS numbers: 43.60.Dh, 43.72.Ja

I. 서 론

무 음성 대화 기술(Silent Speech Interface, SSI)[1]은 실제 발성을 하지 않고 발성 동작만으로 대화하는 기술을 의미한다. SSI는 발성과 연관된 장기(후두)의 영구적 손상으로 음성을 통한 의사전달이 불가능한 발성 장애자를 위한 발성 보조 장치의 개발, 전장과 같은 주변 소음으로 인해 대화가 불가능한 특수한 상황에서 효과적인 의사전달 도구로 이용될 수 있

†**Corresponding author:** Ki-Seung Lee (kseung@konkuk.ac.kr)
Department of Electronic Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea
(Tel: 82-2-450-3489, Fax: 82-2-450-3437)

다. 또한 도서관과 같은 공공장소에서 다른 사람에게 폐를 끼치지 않고 대화가 가능하며, 음성을 통한 정보 누설이 우려되는 상황에서 효과적인 음성은닉 기법으로 활용될 수 있다.

SSI는 기본적으로 비 음성 신호로부터 특징 변수를 추출하고 이로부터 해당 음성의 특징 변수를 추정하는 방법을 통해 구현된다. SSI에 사용되는 비 음성 신호는 가청 주파수 대역의 간섭 신호 영향을 적게 받으며 음성 신호와 높은 상관성을 갖아야 한다.[1] 이러한 조건을 만족하는 신호로서 입 주변에서 취득한 영상 신호,[2] 발성 기관에 대한 초음파 영상 신호,[3] GHz 초음파 신호,[4] 초음파 도플러 신호,[5] 접촉식 마이크로폰으로부터 취득된 신호,[6] 입주변 근육의 근전도 신호[7] 등을 들 수 있다. 이 중 초음파 도플러 신호는 다른 신호에 비해 소형, 경량, 저비용의 장점을 갖는다.[5] 이러한 장점으로, 저자 등은 한국어 음성에 대한 인식 및 합성에 대한 연구를 수행하였으며 다른 비 음성 신호에 비해 우수한 성능을 나타낸 것으로 보고하였다.[5]

초음파 도플러 기반 기법은 발성하고 있는 입 주변에 고정 주파수의 초음파 신호를 방사할 때, 반사파에서 관찰되는 도플러 변이를 검출하여 음성을 합성한다. 이때 관찰되는 도플러 변이는 발성과 연관된 근육의 움직임에 따라 특이적으로 나타내게 되며, 음성 신호와 높은 연관성을 갖게 된다. 기존의 초음파 도플러 기반 기법에서는 도플러 변이가 관찰되는 주파수 대역에서 특징 변수를 추출하고, 해당 음성 신호의 스펙트럼을 추정하도록 하였다.[5] 특징 변수와 음성 스펙트럼 간의 대응관계는 지도학습 기반 기법을 이용하여, 다층 퍼셉트론(Multi-Layer Perceptron, MLP)을 통해 표현하였다. 이때 MLP의 가중치는 최상위 노드에서의 추정값과 실제값 간의 자승오차가 최소화되도록 학습된다. 이와 같은 최소 자승오차 기반 방법은 생성된 음성 파라메터가 실제 파라메터와 유사하게 얻어지며 결과적으로 본래의 음성과 유사한 합성음을 얻지만 분포와 같은 통계적인 특성은 고려되지 않아 작은 오차가 얻어짐에도 생성된 음성의 주관적 품질이 낮아질 수 있다는 문제가 있다. 생성적 적대 신경망(Generative Adversarial Networks, GAN)[8]은 생성기와 함께 분류기를 함께 경쟁적으로 학습시켜 추정된 파라메터의 분포가 실제 파라메터의 분포와 유사하도록 학습하게 된다. 본 논문에서는 기존의 MLP방식의 초음파 도플러 기반 SSI기법에서 GAN을 도입하여 객관적, 주관적인 지표를 통해 성능 향상 여부를 확인하고자 한다.

II. 음향 도플러 센서

Fig. 1과 Fig. 2에 제작된 초음파센서의 사진을 제시하였다. 센서는 40 kHz 정현파 신호를 발생하기 위한 초음파 스피커(Model: MA40H1S-R, Murata)와 반사된 초음파 신호를 취득하기 위한 초음파 마이크로폰(Model: SPM0404UD5, Knowles acoustics)로 구성되어 있으며, 헤드셋 마이크로폰(Model: SHURE BETA 53)의 프레임에 Fig. 1과 같이 장착하였다. 사용된 초음파 마이크로폰의 주파수 응답은 10 kHz ~ 65 kHz로서 40 kHz에 인접된 도플러 변이를 검출하기 용이하나, 음성 대역(~3.5 kHz)의 신호를 취득하는데는 부족하다고 판단되어 헤드셋 마이크로폰으로부터 음성 신호를 취득하였다. 40 kHz 정현파 신호는 함수발

Fig. 1. Photograph of the ultrasonic Doppler sensor, mounted on the right cheek of the subject.

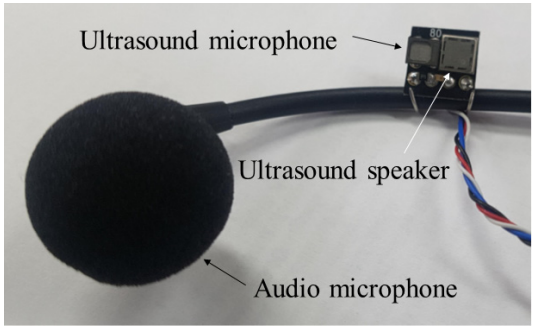

Fig. 2. (Color available online) Photography of the magnified sensor portion.

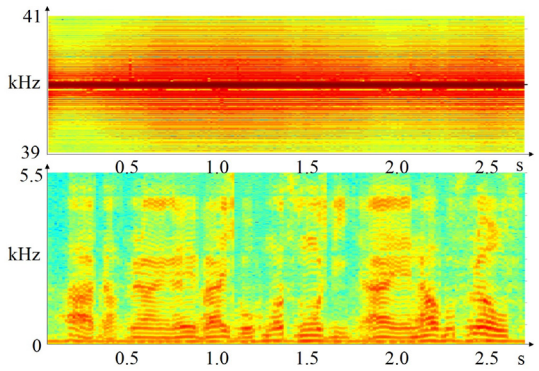

Fig. 3. (Color available online) An example of the spectrograms; Top: recorded ultrasonic signal. Bottom: corresponsing speech signal.

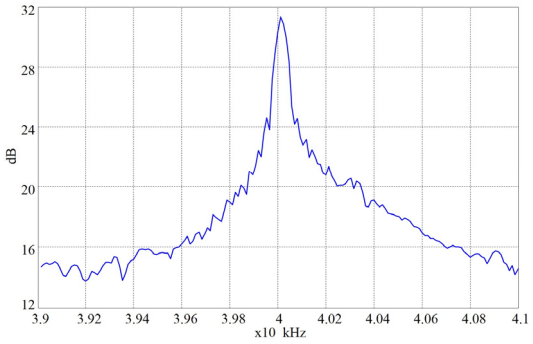

Fig. 4. (Color available online) Power spectral density of the received ultrasonic signals.

생기(Model: 33250A, Agilent)를 통해 생성하고 증폭기(Model: LM386N, National Semiconductor)를 통하여 원하는 신호레벨로 조정하였다. 초음파 스피커의 유효 방사각은 40°이고, 스피커와 피부 간 최소 거리는 3 cm인 조건에서 유효 방사 면적은 19.9 cm^2으로서, 이는 입 주변 발성 움직임을 검출하는데 충분한 면적이었다. 초음파 마이크로폰으로부터 취득된 신호는 디지털 오디오 인터페이스(Model: H6, Zoom)를 통해 샘플링 주파수 96 kHz 해상도 24 bit를 갖는 디지털 값으로 변환하였다. 헤드셋 마이크로폰에서 취득된 가청대역의 음성신호도 동일한 주파수와 해상도로 디지털화하였다.

본 연구와 같이 도플러변이를 이용한 음성 합성 방법은 음성과 무관하게 도플러 변이가 발생한 경우, 합성음의 품질이 크게 저하될 수 있다. 예로서, 심장 박동과 관련된 근육의 움직임이 도플러변이로 나타나게 되면 합성음에 맥박과 일치하는 펄스 잡음이 발생하는 것이 관찰되었다.[5] 본 연구에서 고안된 센서는 헤드셋의 프레임에 장착되어 심장의 박동과 연관된 간섭의 영향을 받지 않으며, 발성과 무관한 얼굴, 목 의 움직임에 따라 음성추정의 정확도가 저하되는 경우가 관찰되지 않았다.

Fig. 3에 제작된 센서를 이용하여 취득된 초음파 신호 및 음성 신호의 스펙트로그램을 제시하였다. 초음파 신호는 입사 신호의 주파수인 40 kHz를 중심으로 음성이 존재하는 구간에서 주파수 변이가 존재함을 알 수 있다. 이러한 주파수 변이는 가청 대역 잡음이 존재하는 경우 잡음의 영향을 받지 않는 것으로 관찰되었으며, 따라서 초음파 도플러를 이용한 음성합성 방법은 배경 잡음에 대해 강인성을 갖는다고 말할 수 있다.[5]

Fig. 4는 취득된 초음파 반사 신호의 전력밀도 함수를 주파수 대역 39 kHz ~ 41 kHz에 대해 도시한 것이다. 취득된 초음파 신호가 실수값으로 주어짐에도 중심 주파수 40 kHz를 기준으로 좌, 우가 약간 다른 형태로 나타난다. 이에 대한 원인으로, 입사 신호의 주파수가 40 kHz를 기준으로 약간 변동되는데, 사용된 함수 발생기가 매우 안정적인 정현파 신호를 발생하지만 사용된 초음파 스피커는 비선형왜곡이 발생하기 때문이다. 또한 전력밀도 함수의 계산 시 사용된 이산 푸리어 변환(Discrete Fourier transform)이 제한된 주파수 해상도를 갖기 때문에 이와 같은 비대칭 형태가 나타난다.

전력밀도 스펙트럼을 관찰하면 40 kHz를 중심으로 매우 좁은 대역에 신호가 분포하는데, 도플러 변이가 음성의 발성에 의한 것임에도 음성 신호의 대역인 3.5 kHz과 비교하여 매우 좁은 대역내에서만 나타난다고 볼 수 있다 . 이는 반사된 초음파 신호의 도플러 변이가 발성과 동반된 피부의 진동 보다는 발성 동작에 의한 것으로 해석할 수 있다. 입 주변 근육에서 취득된 피부 근전도 신호에 대해 주파수 분석을 수행하면 수100 Hz의 낮은 대역에 대해서 신호가 집중적으로 나타나는데,[7] 이는 발성과 연관된 입의 움직임이 낮은 주파수 대역에 분포함을 의미한다. Fig. 4에 제시된 초음파 신호의 전력 밀도 함수는 근전도 신호의 전력 밀도 함수와 대체적으로 일치하며, 이는 초음파 기반 SSI방법이 입주변 근전도 신호

를 사용하는 방법과 유사한 방법임을 의미한다. 그러나 초음파 기반 방법은 근전도 기법과는 달리 비접촉식 센서를 사용하므로 사용자에서 불편감이 덜하다는 장점이 있다.

III. 음성 신호 추정

음성 신호 추정을 위한 특징 변수는 Fig. 4에 제시한 전력 밀도 함수의 분포 형태를 고려하여 얻었다. Fig. 5에 그 과정을 설명하였다. 초음파 도플러 변이는 입사 주파수인 40 kHz 근방에서 나타나므로, 먼저 취득된 초음파 신호를 낮은 주파수 대역으로 이동하기 위한 복조과정을 거친다. 반사 초음파 신호는 Fig. 4에 제시된 바와 같이 40 kHz를 중심으로 좌, 우가 약간 다르게 나타나므로 복조 시 주파수 이동은 39 kHz로 설정하였다. 이 경우, 중심 주파수는 1 kH가 되고 좌/우 각 1 kHz의 폭을 갖게 된다.

초음파 도플러 변이는 발성과 동반된 피부의 진동이 아닌, 발성과 연관된 근육의 움직임에 의해 발생하였기 때문에 복조된 스펙트럼은 음성 신호의 스펙트럼과는 다른 형태를 갖게 된다. 그러나 근전도 신호를 이용한 음성 합성 및 초음파 도플러를 이용한 음성 합성 연구 결과에 따르면, 멜-주파수 필터 뱅크를 이용하여 대역 분할된 신호를 입력 변수로 사용

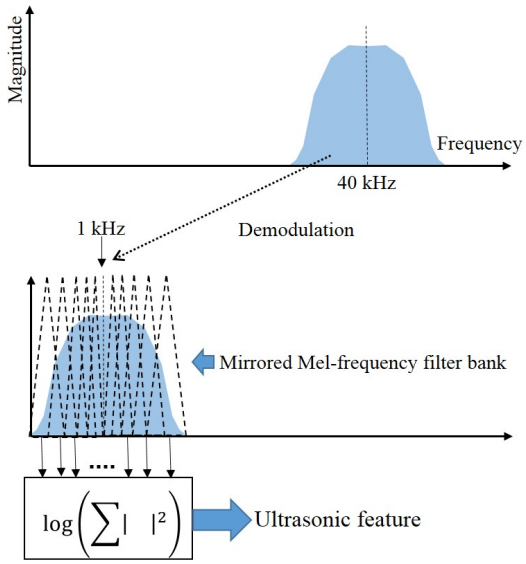

Fig. 5. (Color available online) Extraction of the ultrasonic Doppler feature parameters.

한 경우 우수한 성능을 나타내는 것으로 보고하였다.[5,7] 이러한 연구 결과에 따라, 본 논문에서도 복조된 스펙트럼에 대해 멜-주파수 필터 뱅크를 통과한 각 신호의 크기를 특징 변수로 사용하였다. 특징 변수의 개수는 좌, 우 8개의 대역통과 필터를 사용하여 총 16개로 설정하였다. 특징 변수의 추정은 취득된 초음파 신호에 대해 창 함수를 곱하여 단 구간으로 분할하고 각 단 구간 신호(프레임)별로 계산된다. 이 때 frame의 이동 간격은 음성 신호의 합성 간격과 동일하게 설정하였다.

단 구간 분석된 음성 신호는 인접 프레임 간에 높은 상관성을 갖는 것으로 알려져 있다. 초음파 도플러 신호 역시 음성으로부터 유래된 신호이므로 인접된 특징 변수 간에 높은 상관성을 갖는 것으로 가정할 수 있다. 따라서 본 논문에서는 n-번째 프레임에 대한 음성 신호 추정 시 좌, 우 각 2개의 프레임에서 계산된 초음파 특징 변수를 함께 사용하였다. 따라서, 추정기의 입력 신호는 다음과 같이 나타낼 수 있다.

$$\mathbf{X}_n = [\mathbf{x}_{n-2}, \mathbf{x}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n+1}, \mathbf{x}_{n+2}], \quad (1)$$

여기서 $\mathbf{X}_n$은 n-번째 프레임에 대한 음성 추정을 위한 특징 변수를 나타내며 $\mathbf{x}_n$은 n-번째 프레임에 대한 멜-주파수 대역 통과 신호를 나타낸다.

본 논문에서는 음성 추정 시 신경 회로망을 이용한 지도 학습 기법이 사용되었다. 신경 회로망의 학습에는 역 전파 알고리즘이 사용되었으며 목적 함수로 아래 식으로 주어지는 평균 자승 오차(Mean Square Error, MSE)가 사용되었다.

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{p=1}^{P}\{F_p(\mathbf{W}, \mathbf{X}_n) - \mathrm{y}_{p,n}\}^2, \quad (2)$$

여기서 $F_p(\mathbf{W}, \mathbf{X}_n)$은 가중치가 $\mathbf{W}$로 주어지는 신경망의 $p-$번째 출력을 나타내고 $\mathrm{y}_{p,n}$는 $n-$번째 프레임에 대한 $p-$번째 주파수 성분을 나타낸다. 본 연구에서 $\mathrm{y}_{p,n}$는 단구간 음성 신호에 대해 이산 푸리어 변환을 적용하고, 푸리어 계수값에 log- magnitude를 취함으로써 얻어진다. N과 P는 각각 학습 데이터의 전체 개수와 푸리어 계수의 개수를 나타낸다. 8 kHz로

샘플링된 음성 신호에 대해 본 연구에서는 $P=128$로 설정하였다. 가중치는 MSE값에 대한 gradient를 취하여 반복적으로 갱신하였다.

$$\mathbf{W}_{n+1}=\mathbf{W}_n-\lambda\nabla_{\mathbf{W}}E, \tag{3}$$

여기서 λ는 학습 레이트를 나타내며 본 연구에서는 0.0001값을 사용하였다.

MSE를 최소화하는 관점에서 학습된 신경망은 실제 음성 스펙트럼과 유사한 출력 값을 얻을 수 있으나, 통계적으로 실제 음성과 유사한 특성을 보장할 수 없다. 생성적 적대 신경망(GAN)[8]은 이와 같은 기존 신경망 기반 추정기의 문제점을 해결하기 위한 방법으로서, Fig. 6에 구조를 나타내었다. GAN에서는 두 개의 신경망이 존재하는데, 하나는 생성기로서, 입력 특징 변수로부터 목표 특징 변수와 유사한 변수를 얻는데 사용한다. 기존 신경망 기반 SSI에서 음성 스펙트럼을 추정하는데 사용한 신경망이 생성기에 해당한다. 다른 하나의 신경망은 분류기로서, 추정된 파라메터와 목표 파라메터, 즉, fake와 real 파라메터를 구분하기 위해 도입되었다. 분류기 학습에는 생성기로부터 얻어지는 출력 값(추정 음성 스펙트럼)과 실제 값(실제 음성 스펙트럼) 각각에 대해 '0'과 '1'로 레이블링 된 data를 사용한다. 분류기의 학습은 loss 값 $L_D(\mathbf{Y},\hat{\mathbf{Y}})$를 최소화하는 방향으로 이루어진다.

$$L_D(\mathbf{Y},\hat{\mathbf{Y}})=-\frac{1}{N}\sum_{n=1}^{N}\log\frac{1}{1+\exp\{-D(\mathrm{Y}_n)\}}-\frac{1}{N}\sum_{n=1}^{N}\log\left(1-\frac{1}{1+\exp\{-D(\hat{\mathrm{Y}}_n)\}}\right), \tag{4}$$

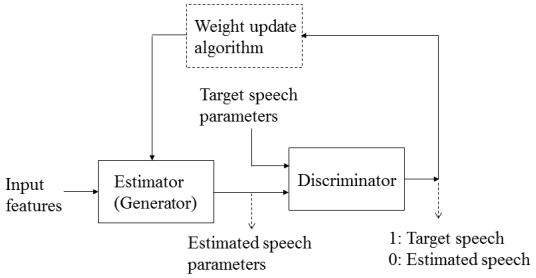

Fig. 6. Block diagram of generative adversarial networks.

여기서 $D(Y)$은 분류기에 X가 입력되었을 때 출력 값을 나타낸다.

생성기의 학습에는 학습된 분류기의 출력이 '1'이 되도록, 즉 추정된 음성 스펙트럼이 실제 음성 스펙트럼으로 판정되도록, 아래 식으로 주어지는 loss값 $L_G(\hat{\mathbf{Y}}=F(\mathbf{W},\mathbf{X}))$을 최소화하는 방향으로 이루어진다.

$$L_G(\hat{\mathbf{Y}})=\frac{1}{N}\sum_{n=1}^{N}\log\frac{1}{1+\exp\{-D(\hat{\mathrm{Y}}_n)\}}. \tag{5}$$

이와 같은 방법을 통해 학습된 생성기는 추정된 음성 스펙트럼의 분포가 실제 음성 스펙트럼의 분포와 유사하게 얻어지지만, MSE를 loss 함수로 사용한 기존의 방법과 비교하여 스펙트럼의 형태면에서 유사성이 저하될 수 있다. 따라서 본 논문에서는 기존 MSE와 Eq. (4)로 주어지는 GAN-loss 값을 선형 조합하여 생성기를 학습하는 방법을 제안하였다. 제안된 기법의 생성기 loss 함수는 다음과 같다.

$$L_G'=(1-\omega_G)E+\omega_G\frac{\mu_E}{\mu_G}L_G, \tag{6}$$

여기서 μ_E와 μ_G는 각각 MSE loss와 GAN-loss에 대한 평균을 나타낸다. ω_G는 GAN-loss에 대한 상대적인 가중치를 나타내며, $0\le\omega_G\le1$의 조건을 만족한다. MSE-loss에 대한 상대적인 가중치는 $1-\omega_G$로 주어진다. 제안된 음성 추정기법의 블록도를 Fig. 7에 제시하였다. 이와 같은 방법을 통해 학습된 생성기는 실제 음성 스펙트럼의 형태 및 분포가 유사하도록 음성 스펙트럼의 추정이 가능하다. 또한 상대적인 가중치를 적절히 조절하여 스펙트럼 형태 우선,

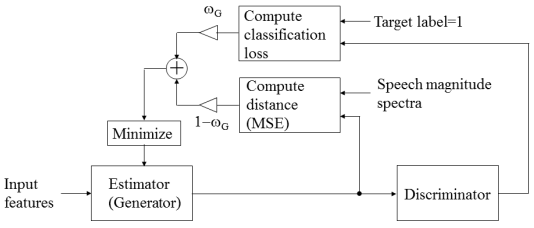

Fig. 7. Block diagram of the proposed speech estimation scheme.

또는 분포 우선의 학습 방법이 구현 가능하며, 이에 대한 결과는 다음 장에서 제시한다.

IV. 실험 및 결과

제안 기법의 유효성을 검증하기 위해 한국어 60단어[7]에 대해 음성을 합성하고 다른 기법과 비교하였다. 실험 데이터는 3명의 피시험자(20대 남성2, 20대 여성1)로부터 비교적 조용한 환경에서 60개 단어를 30차례 반복 녹음하여 이 중 20개는 신경망 학습에, 10개는 신경망의 학습에 필요한 파라메터(학습 레이트, 미니배치 크기, 드롭아웃 확율) 결정에 사용하고, 나머지 10개는 test데이터로 사용하였다. 피시험자 간 발성 스타일의 차이로 인해 초음파 도플러 패턴은 피시험자 별로 상이하게 나타났으며, 따라서 음성 추정 규칙은 피시험자별로 개별적으로 생성하였다. 신경망에 입력되는 초음파 신호의 특징 변수는 250 msec의 길이를 갖는 hamming창 함수를 100 msec 만큼 이동하면서 추출하였다. 음성 스펙트럼의 표현에 필요한 푸리어 변환 계수의 추출에도 동일한 조건이 적용되었다.

본 논문에서는 음성 파라메터로서 Fourier transform magnitude spectrum을 사용하기 때문에 음성을 합성하기 위해서는 추가적으로 phase spectrum에 대한 정보가 필요하다. 본 연구에서는 magnitude spectrum으로부터 phase spectrum을 추정하기 위한 기법으로 Griffin-Lim이 제안한 반복 추정 방법[9]을 사용하였다.

신경망의 구조는 수차례의 실험을 통해 결정하였는데, 최종적으로 생성기와 분류기 모두 3개의 은닉 계층을 갖고, 각 은닉 계층 마다 240개의 노드를 갖는 구조로 결정하였다. Sigmoid active함수가 사용되었으며 출력노드의 개수는 128로 설정하였다. Batch size, dropout probability, learning rate는 각각 10, 0.75, 0.0001로 결정하였다.

성능 평가를 위한 지표로서 본 논문에서는 합성음에 대한 Perceptual Evaluation of Speech Quality(PESQ),[10] 추정된 음성 스펙트럼과 실제 스펙트럼간의 Root Mean Square Error(RMSE)를 사용하였고, 분류기에 대해서는 인식율을 사용하였다. 제안된 기법과의 성능 비교를 위해 기존 MSE만을 사용한 초음파 도플러 기반 SSI방법을 결과를 제시하였다.

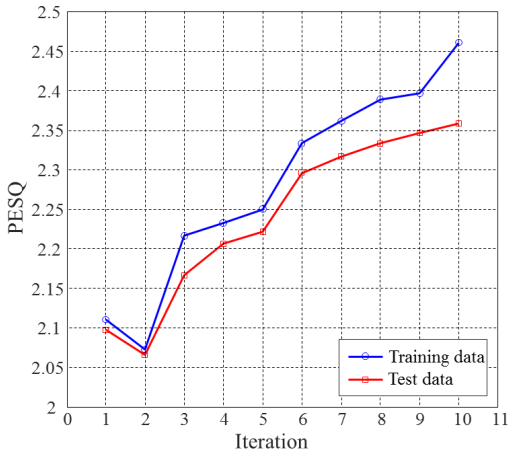

Fig. 8. (Color available online) PESQs vs. Iteration for training/test data.

실험결과는 먼저 생성기와 분류기가 서로 반복적으로 학습하는 과정에서 PESQ가 어떻게 변동되는지를 관찰하였다. 만약 epoch가 증가함에 따라 PESQ가 증가되는 것이 확인되면 분류기와 생성기가 주어진 loss함수를 지속적으로 감소시키고, 이는 본래의 음성과 합성된 음성이 청감상으로 점차 가까워짐을 의미한다. 반대로 epoch가 증가하더라도 PESQ의 변화가 없다면 비록 loss함수가 감소하더라도 음질적인 향상과는 무관하게 학습이 진행되는 것으로 해석할 수 있다. 이에 대한 결과를 Fig. 8에 제시하였다. 그림의 결과는 초기 epoch 1 ~ 10에서 추정된 음성 스펙트럼를 이용하여 합성된 음성의 PESQ를 도시한 것이다. 그림에서 보듯이, epoch가 증가함에 따라 PESQ가 지속적으로 증가하는 것을 알 수 있으며, 학습 데이터와 테스트 데이터에게 공통적으로 나타남을 알 수 있다. 다만 epoch 1 → 2에서는 일시적으로 PESQ가 감소되는 것을 알 수 있다. 이러한 일시적인 감소는 learning rate를 epoch에 따라 적절히 조절하거나 loss함수의 감소율을 고려하여 신경망의 가중치를 조정하는 방법으로 해결 가능하다. Fig. 8의 결과는 제안된 학습 방법에서 사용된 loss함수는 PESQ를 증가시키는데 효과적임을 알 수 있다.

두 번째로 분류기가 각 epoch별로 어떠한 정확도를 보이는 가 관찰하였다. 추정 스펙트럼이 실제 음성 스펙트럼의 분포와 유사하게 얻어지기 위해서는

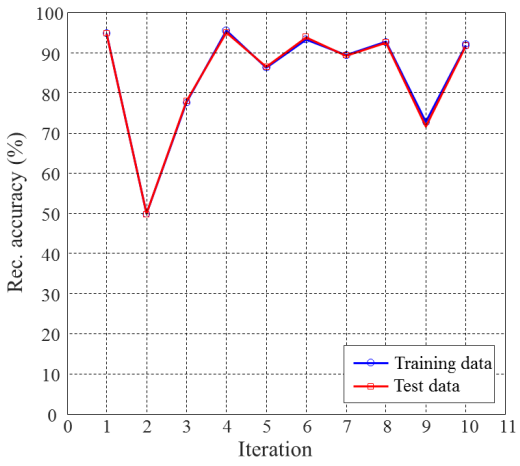

Fig. 9. (Color available online) Recognition accuary vs. Iteration for training/test data.

높은 인식율을 갖는 분류기가 필수적이기 때문이다. epoch에 따른 인식율을 Fig. 9에 도시하였다. 그림을 보면 epoch 1→2에서 급격히 감소하고, 이후 90 %를 상회하는 높은 인식율을 보이다가 epoch 9에서 다시 감소하다가 증가하는 양상을 보이고 있다. 이는 epoch 2에서 추정된 음성 스펙트럼이 실제 음성 스펙트럼과 유사도가 증가하고 이에 따라 두 스펙트럼 간의 모호성이 일시적으로 증가하면서 저하된 분류 성능을 보이는 것으로 해석할 수 있다. 이후, 분류기가 반복적으로 학습되면서 90 %에 가까운 안정된 인식율을 보인 후 epoch 9에서 모호성이 다시 증가하면서 인식율이 저하되는 것으로 보인다. 이러한 양상은 epoch-10 이 후에도 반복적으로 관찰되었는데, 생성기에서 추정된 스펙트럼이 점진적으로 실제 음성 스펙트럼과 가까워지면서 모호성이 증가되고, 모호성이 임계치를 초과하게 되면 분류기의 성능이 급격히 저하되고 이 후 다시 분류기가 학습되면서 안정된 인식율을 보이는 패턴이 반복적으로 나타난다고 볼 수 있다.

Table 1에 3명의 피시험자에 대한 PESQ 및 RMSE 값을 가중치 ω_G 별로 제시하였다. 표에서 ω_G=0에 해당하는 결과는 GAN-loss를 고려하지 않은 경우로서, 기존의 MSE기반 음성 추정 방법과 동일한 결과라고 간주 할 수 있다. 실험 결과를 살펴보면 모든 피시험자에 대해 $0 < \omega_G \leq 0.5$인 경우, MSE만을 사용한 기존의 방법에 비해 높은 PESQ를 나타내었다. 반면,

Table 1. PESQs (top) and RMSEs (bottom) for the subjects according to the weight value for MSE.

	0.0	0.25	0.5	0.75	1.0
subject1	2.225	2.324	2.366	2.211	2.205
subject2	2.237	2.409	2.409	2.265	2.201
subject3	2.321	2.362	2.421	2.283	2.147
	0.0	0.25	0.5	0.75	1.0
subject1	0.0285	0.0291	0.0295	0.0310	0.0321
subject2	0.0275	0.0278	0.0287	0.0292	0.0301
subject3	0.0277	0.0279	0.0281	0.0284	0.0286

RMSE값은 $\omega_G > 0$인 모든 경우, MSE만을 고려한 경우와 비교하여 높게 나타났는데, 이는 제안방법이 MSE뿐이 아니고 GAN-loss도 함께 최소화하기 때문이다. 결론적으로 제안된 신경망 학습 방법은 MSE와 GAN-loss를 함께 고려하여 최소화를 수행하며, 각 loss값에 대한 가중치를 적절히 조정함으로써 기존의 MSE만을 고려한 방법과 비교하여 음질적으로 우수한 음성을 합성한다고 볼 수 있다. 실험에서는 모든 피시험자에 대해 가중치가 0.5인 경우, 즉 MSE를 GAN-loss 동등하게 조합한 경우 최대 PESQ값이 얻어짐을 알 수 있었다. 가중치가 0.5를 초과한 경우는 MSE의 비중이 상대적으로 낮아지면서 추정된 스펙트럼과 실제 스펙트럼 간의 오차가 과도하게 증가하여 PESQ가 감소한 것으로 판단된다.

PESQ는 청감상 두 음성 간의 차이를 잘 나타내는 척도이지만, 본 논문에서는 청취자 관점에서 합성음의 품질을 주관적으로 평가하기 위해 선호도 테스트를 수행하였다. 선호도 테스트는 두 가지 방법(MLP 방법 $\omega_G = 0$, GAN방법 $\omega_G = 0.5$)으로 합성된 음성을 무작위로 선택하여 10명의 청취자에게 헤드폰을 이용하여 들려주고 어떠한 음성이 선호되는지를 조사하는 방식으로 진행하였다. 다중 선택은 허용하지 않았고, 판단이 이루어지기 까지 반복하여 청취하는 것을 허용하였다. 각 피시험자의 음성에 대한 선호도 결과를 Fig. 10에 제시하였다. 모든 피시험자의 음성에 대해 GAN방법으로 합성된 음성이 높은 선호도를 나타내었다. 이는 Table 1에 제시된 PESQ결과와 일치하는 것으로 GAN을 이용한 음성 추정 방법이 기존의 신경망 방법보다 실제 청취 시에도 음질적인 향상을 가져옴을 입증하는 것이다.

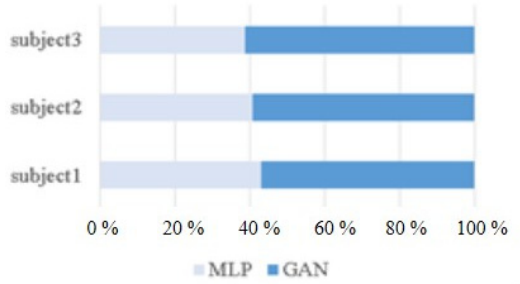

Fig. 10. (Color available online) Preference test results.

V. 결 론

SSI에서 음성 파라메터를 추정하는 방법으로 기존의 MSE만을 최소화하는 학습 방법에 비해 추정 파라메터의 분포를 함께 고려한 방법은 주관적, 객관적으로 우수한 성능을 나타냄을 확인할 수 있었다. 본 논문에서는 MSE, GAN-loss 두 개의 loss값을 사용하고 있는데, 청감 상 차이를 반영한 perceptual distance를 추가하고, 음성의 fine structure를 고려한 추정 방법이 고안되면 보다 더 우수한 음질을 얻을 수 있을 것으로 기대된다.

감사의 글

본 논문은 한국연구재단의(연구과제명: 극심한 소음환경에서 보조 신호를 사용한 음성 복원, S2020 03S00170) 지원을 통해 수행된 연구 결과입니다.

References

1. B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, “Silent speech interfaces,” Speech Comm. **52**, 270-287 (2010).
2. T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone, “Eigentongue feature extraction for an ultrasound-based silent speech interface,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 1245-1248 (2007).
3. B. Denby and M. Stone, “Speech synthesis from real time ultrasound images of the tongue,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 685-688 (2004).
4. S. Li, Y. Tian, G. Lu, Y. Zhang, H. Lv, X. Yu, H. Xue, H. Zhang, J. Wang, and X. Jing, “A 94-GHz milimeter-wave sensor for speech signal acquisition,” Sensors, **13**, 14248-14260 (2013).
5. K. S. Lee, “Silent speech interface using Doppler sonar,” IEICE Trans. on Information and Systems, **E103-D**, 1875-1887, (2020).
6. T. Toda and K. Shikano, “NAM-to-Speech convertsion with Gaussian Mixture Models,” Proc. INTERSPEECH, 1957-1960 (2005).
7. K.-S. Lee, “Prediction of acoustic feature parameters using myoelectric signals,” IEEE Trans. on Biomed. Eng, **57**, 1587-1595 (2010).
8. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Proc. Adv. NIPS. 2672-2680 (2014).
9. D. W. Griffin and J. S. Lim, “Signal estimation from the modified short-time fourier transform,” IEEE Trans. on ASSP. **32**, 236-243 (1984).
10. ITU-T, Rec. P. 862, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for end-to-end Speech Quality Assessment of Narrow Band Telephone Networks and Speech Codecs*, 2001.

저자 약력

▸ 이 기 승 (Ki-Seung Lee)

1991년 2월 : 연세대 전자공학과 학사
1993년 2월 : 연세대 전자공학과 석사
1997년 2월 : 연세대 전자공학과 박사
2000년 9월 : AT&T Shannon Labs
2001년 8월 : 삼성전자(주) 종합기술원
2001년 9월 ~ 현재 : 건국대학교 전기전자 공학부 교수