한국음향학회지 제38권 제4호 pp. 433~441 (2019)
The Journal of the Acoustical Society of Korea Vol.38, No.4 (2019)
https://doi.org/10.7776/ASK.2019.38.4.433

pISSN : 1225-4428
eISSN : 2287-3775

LSTM 순환 신경망을 이용한 초음파 도플러 신호의 음성 패러미터 추정

Estimating speech parameters for ultrasonic Doppler signal using LSTM recurrent neural networks

주형길,[1†] 이기승[1]

(Hyeong-Kil Joo[1 †] and Ki-Seung Lee[1])

[1]건국대학교 전기전자공학부

(Received May 15, 2019; revised July 3, 2019; accepted July 11, 2019)

초 록: 본 논문에서는 입 주변에 방사한 초음파 신호가 반사되어 돌아올 때 발생하는 초음파 도플러 신호를 LSTM (Long Short Term Memory) 순환 신경망 (Recurrent Neural Networks, RNN)을 이용해 음성 패러미터를 추정하는 방법을 소개하고 다층 퍼셉트론 (Multi-Layer Perceptrons, MLP) 신경망을 이용한 방법과 성능 비교를 하였다. 본 논문에서는 LSTM 순환 신경망을 이용해 초음파 도플러 신호로부터 음성 신호의 푸리에 변환 계수를 추정하였다. LSTM 순환 신경망을 학습하기 위한 입력 및 기준값으로 초음파 도플러 신호와 음성 신호로부터 각각 추출된 멜 주파수 대역별 에너지 로그값과 푸리에 변환 계수가 사용되었다. 테스트 데이터를 이용한 실험을 통해 LSTM 순환 신경망과 MLP의 성능을 평가, 비교하였고 척도로는 평균 제곱근 오차(Root Mean Squared Error, RMSE)가 사용되었다. 각 실험의 RMSE는 각각 0.5810, 0.7380로 나타났다. 약 0.1570 차이로 LSTM 순환 신경망을 이용한 방법의 성능 우세한 것으로 확인되었다.

핵심용어: 무음성 인터페이스, 음성 패러미터, 초음파 도플러 신호, LSTM (Long Short Term Memory) 순환 신경망

ABSTRACT: In this paper, a method of estimating speech parameters for ultrasonic Doppler signals reflected from the articulatory muscles using LSTM (Long Short Term Memory) RNN (Recurrent Neural Networks) was introduced and compared with the method using MLP (Multi-Layer Perceptrons). LSTM RNN were used to estimate the Fourier transform coefficients of speech signals from the ultrasonic Doppler signals. The log energy value of the Mel frequency band and the Fourier transform coefficients, which were extracted respectively from the ultrasonic Doppler signal and the speech signal, were used as the input and reference for training LSTM RNN. The performance of LSTM RNN and MLP was evaluated and compared by experiments using test data, and the RMSE (Root Mean Squared Error) was used as a measure. The RMSE of each experiment was 0.5810 and 0.7380, respectively. The difference was about 0.1570, so that it confirmed that the performance of the method using the LSTM RNN was better.

Keywords: Silence speech interface, Speech parameters, Ultrasonic Doppler signal, LSTM (Long Short Term Memory) RNN (Recurrent Neural Networks)

PACS numbers: 43.72.Bs, 43.72.Kb

I. 서 론

인간은 음성을 사용한 의사 전달 방법을 가장 많이 이용한다. 음성은 화자로부터 발성이 되고 공기를 매질 삼아 청자에게 전달이 된다. 그런데 음성 전

†**Corresponding author:** Hyeong-Kil Joo (mrjoohk@gmail.com)
Department of Electronic Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea
(Tel: 82-2-450-3489, Fax: 82-2-3437-5235)

달은 유동인구가 많은 길거리나 열차 안과 같이 주변 소음이 매우 큰 주변 환경에 의해서 제한되는 경우가 있다. 또한, 도서관이나 영화관 같은 장소에서 서로를 위해 대화에 의한 큰 소리를 내지 않아야 하는 경우나 대화 내용이 다른 사람에게 전달되지 않아야하는 특수한 경우처럼 화자 스스로가 음성 전달을 제한해야 하는 경우가 있다. 무음성 인터페이스 방법[1]은 이처럼 음성 전달을 제한해야 하는 경우를 극복하기 위한 의사전달 방법으로 제안되었다. 이 방법은 음성을 내기 위한 일련의 과정으로부터 소리가 아닌 다른 신호를 취득해 이를 이용하여 의사전달을 가능하게 한다. 기존에 연구된 무음성 인터페이스 방법을 구현하는 방법으로는 입 주변에 부착한 전극으로부터 취득한 근전도 신호를 이용하여 음성신호를 추정하는 방법[2], 귀 후면에 부착한 마이크로폰을 이용하여 화자의 웅얼거림에서 발생하는 미세한 진동을 취득해 음성신호를 추정하는 NAM(Non-Audible Murmur) 방법,[3] 입 주변에 마이크로파를 발사하고, 반사되는 신호의 도플러를 이용한 방법,[4] 그리고 초음파 도플러를 이용한 방법[5] 등이 있다.

본 논문에서는 초음파 도플러를 이용한 방법으로 무음성 인터페이스를 구현하였다. 초음파 도플러는 파원에서 발사한 초음파가 움직이는 물체에 닿아 반사될 때 파원과 물체 사이의 상대속도 변이에 따라 반사되는 초음파의 주파수가 본래와는 다른 주파수를 갖게 되는 현상이다. 초음파 도플러를 이용한 연구는 Kalgaonkar *et al*.,[6-8] Srinivasan *et al*.[9] 그리고 Livescu *et al*.[10] 에 의해 이미 진행되었고 효용성을 확인하였다. 그리고 Toth *et al*.[11]의 연구에서는 혼합 가우시안 모델(Gaussian Mixture Model, GMM)을 이용하여 초음파 도플러 신호를 음성신호로 변환하여 청취 시 인지 가능한 합성음을 얻을 수 있음을 보여주었다. 국내에서도 Lee[12-14]의 연구에서 한국어에 대해 초음파 도플러 기반 음성합성의 가능성을 제시하였다. Reference [14]에서는 인공 신경망의 일종인 다층 퍼셉트론(Multi-Layer Perceptrons, MLP) 신경망을 이용하여 음성 합성을 시도하였고 기존 방법들보다 향상된 인지율을 보여주었다. 그러나 다층 퍼셉트론 신경망은 입력과 출력 사이의 일대일 관계만을 이용해 학습하고 시계열 데이터의 특성을 고려하지 않아 성능의 한계가 있다. 이에 따라 본 논문에서는 과거에 입력받았던 데이터와 현재 입력 데이터로 다음 데이터를 추정하는 인공 신경망의 일종인 LSTM (Long Short Term Memory) 순환 신경망(Recurrent Neural Networks, RNN)을 사용하여 초음파 도플러 신호로부터 음성 합성을 위한 음성 패러미터를 추정하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 초음파 도플러 신호에 대한 분석, 초음파 도플러 신호로부터 특징변수를 추출하는 방법, 다층 퍼셉트론 신경망 그리고 본 논문에서 제안하는 LSTM 순환 신경망에 대해 설명한다. 그리고 3장에서는 실험 데이터 취득 방법, 제안된 LSTM 순환 신경망과 MLP를 각각 이용한 실험 결과로 성능의 평가와 비교를 하고, 마지막으로 4장에서 결론을 맺는다.

II. 연구 방법

본 논문에서 제안하는 기법은 Fig. 1에서 나타낸 것처럼 학습 단계와 테스트 단계로 나뉜다. 학습 단계에서는 취득된 초음파 도플러 신호로부터 특징변수를 추출하는 과정, 추출된 특징변수와 음성신호로부터 추출된 푸리에 변환(Fourier transform) 계수를 각각 신경망의 입력과 기준값으로 사용한 신경망의 학습 과정이 진행된다. 그리고 테스트 단계에서는 학습 단계에서 학습된 신경망에 테스트 데이터로부터 추출된 특징변수를 입력으로 주어 음성신호 푸리에 변환 계수를 추정하는 과정, 성능 척도에 따른 성능 평가, 비교가 진행된다.

2.1 초음파 도플러 신호 분석 및 특징변수 추출

인간이 음성을 내려면 입과 입 주변 근육의 움직

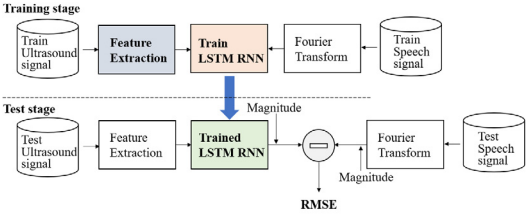

Fig. 1. A block diagram of the proposed method.

임, 그리고 목의 떨림 등과 같은 일련의 과정이 동반되어야 한다. 따라서 음성을 낼 때, 입 주변에 초음파를 방사하면 초음파가 부딪히는 반사면인 입 주변은 움직이게 되고 이로 인해 도플러 현상이 발생해 주파수가 변이된 신호, 즉 도플러 신호가 반사된다. 방사된 초음파 신호 $T(t)$를 주파수 f_s, 크기 A_s, 위상 θ_s를 가지는 정현파 신호라 한다면 다음 Eq. (1)과 같이 나타낼 수 있다.

$$T(t)=A_s\cos\left(2\pi f_s t+\theta_s\right). \tag{1}$$

그리고 아래 Eqs. (2)~(4)에서 나타낸 것처럼 입 주변의 N개의 반사면으로부터 발생하는 도플러 신호들을 각각 초음파의 속도 v_s, 센서의 관점에서 반사면의 상대속도 $v_i(t)$에 대해 변이된 주파수 $f_i(t)$, 크기 A_i, 위상 θ_i를 가지는 신호 $d_i(t)$ ($i=1, 2, \cdots, N$)라 한다면 센서에 취득되는 신호 $R(t)$는 N개의 도플러 신호가 중첩된 형태로 다음과 같이 나타낼 수 있다.

$$f_i(t)=f_s\left(\frac{v_s+v_i(t)}{v_s-v_i(t)}\right), \tag{2}$$

$$d_i(t)=A_i\cos\left(2\pi f_i(t)+\theta_i\right), \tag{3}$$

$$R(t)=\sum_{i=1}^{N} d_i(t). \tag{4}$$

취득된 신호 $R(t)$는 초음파 중심 주파수 40 kHz에 대한 2 kHz 대역폭을 지닌 대역통과필터를 통과 후, 중심 주파수를 1 kHz로 하향 변환하는 복조(Demodulation)를 수행하고 특징변수를 추출했다.[9] 초음파 도플러 신호는 음성을 내기 위한 일련의 과정으로부터 발생하므로 음성신호와 유사한 특성을 가지고 있음을 추측할 수 있다. 얼굴 부위 근전도 신호로부터 추출한 특징변수를 사용한 연구[2]와 입 주변에 방사한 초음파가 반사되어 발생한 초음파 도플러 신호로부터 추출한 특징변수를 사용한 연구[12-14]의 높은 인식률을 보인 결과들로부터 음성을 내기 위한 일련의 과정으로부터 발생한 신호와 음성신호 간에 유사한 특성을 가지고 있음을 확인할 수 있었다. 이에 따라

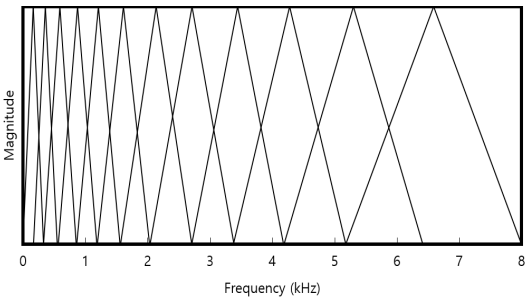

Fig. 2. Band pass characteristics for each mel frequency band.

본 연구에서는 초음파 도플러 신호로부터 추출한 멜 주파수 대역별 에너지 로그값을 특징변수로 사용해 실험을 진행하였다.

멜 주파수 대역별 에너지 로그값은 로가리듬(Logarithm)한 비선형적 주파수 특성을 가지는 인간의 청각 특성이 반영된 특징변수이고 다음과 같은 방법으로 얻을 수 있다. 우선, 주어진 신호를 200 msec 길이의 구형함수(Rectangular function)를 100 msec씩 겹치도록 이동하며 윈도윙(Windowing)해 프레임 단위로 변환 후 프레임들을 푸리에 변환하여 주파수 영역으로 변환한다. 그리고 Fig. 2와 같이 주파수 대역을 나눈 여러 개의 필터뱅크에 대해서 각각 대역 에너지를 구하고 로그를 취해 얻을 수 있다. 초음파 도플러 신호의 M 포인트 푸리에 변환 계수에 대해 m-번째 계수를 $X[m]$, 그리고 i-번째 멜 주파수 대역통과 특성뱅크의 m-번째 계수를 $H_i[m]$이라 했을 때 i-번째 멜 주파수 대역 에너지 로그값 Y_i는 다음 Eq. (5)와 같이 나타낼 수 있다.

$$Y_i=\log\left[\sum_{m=1}^{M}|X[m]|H_i[m]\right]. \tag{5}$$

2.2 MLP

퍼셉트론은 인간의 신경 세포의 수학적 모델이자 최초의 신경망 알고리즘이다. 퍼셉트론 알고리즘은 학습 데이터를 잘 구분할 수 있는 최적의 가중치와 편향치를 찾도록 학습한다.[15] 여러 개의 퍼셉트론을 연결하여 층을 만들고, 이 층들을 여러 층으로 쌓아 올려 다층 퍼셉트론을 만든다. 다층 퍼셉트론은 Fig. 3에서 나타낸 것처럼 입력층, 은닉층, 출력층으로 이

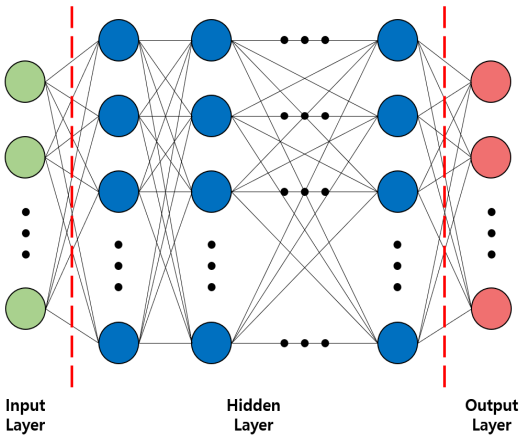

Fig. 3. Structure of MLP.

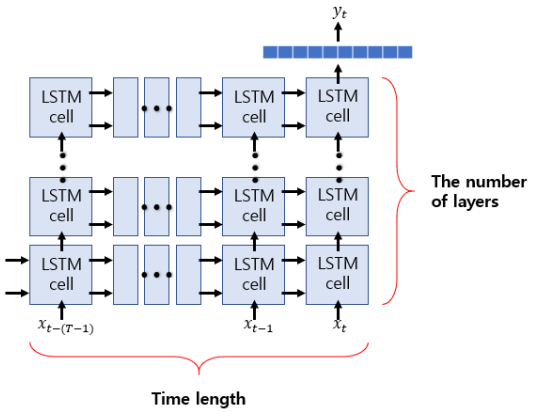

Fig. 4. Structure of LSTM RNN.

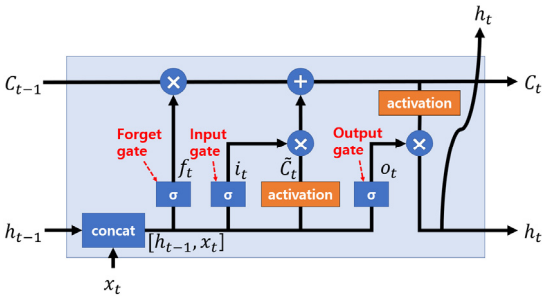

Fig. 5. Structure of LSTM cell.

루어져있으며 역전파 알고리즘을 이용하여 학습이 이뤄진다.[16] 본 연구에서는 2개 ~ 5개의 은닉층과 각 은닉층에 640개 ~ 1920개의 노드를 사용하여 LSTM 순환 신경망과의 성능 비교를 하였다.

2.3 LSTM 순환 신경망

일반적으로 과거의 사건들은 앞으로 일어날 사건들에 영향을 준다. 순환 신경망은 이러한 아이디어에 기반하여 이전 단계의 정보가 다음 단계로 전달되어 추정하는데 사용되는 구조를 가지고 있다. 이에 따라 순환 신경망은 순차적인 정보, 즉 시계열 데이터의 추정에 큰 성과를 거두었다. 그러나 기존의 순환 신경망은 데이터 추정을 위해 요구하는 과거 데이터의 길이가 길어지면 길어질수록 Vanishing Gradient 문제가 발생하고 이로 인해 과거 정보가 현재까지 전달되는 과정에서 소실되어 충분히 전달되지 못해 발생하는 장기 의존성 문제를 가지게 된다. 이 문제점은 Hochreiter & Schmidhuber이 제안한 LSTM 알고리즘의 게이트(Gate)들을 사용함으로써 해결되었다.[17-18] LSTM 순환 신경망은 Fig. 4[19]에서 나타낸 것처럼 가로세로 연결된 셀들과 최상단의 전결합 레이어(Fully-connected layers)구조로 이루어져있다. LSTM 셀의 세로축 길이는 깊은 학습(Deep learning)을 위해 쌓은 LSTM 레이어 수를 의미하고 가로축 길이는 데이터를 추정하는데 필요한 입력의 시간 프레임 개수 T에 의해 결정된다. 시간 인덱스 t에 대해 길이 N의 특징변수 프레임을 x_t라 할 때, T개 셀들은 각각 $x_{t-(T-1)}$, $x_{t-(T-2)}$, ⋯, x_{t-1}, x_t를 입력으로 받고 이전 셀의 출력과 셀 스테이트(Cell state)를 전달받아 업데이트하여 다음 셀에 전달한다. 이전 LSTM 레이어의 출력들은 다음 레이어의 입력으로 전달되고 마지막 LSTM 레이어 시간 t 셀의 출력은 최상단의 전결합 레이어들을 통과하여 최종적으로 추정값이 출력된다.

출력 레이어의 활성함수는 일반적으로 추정 대상이 정성적이라면 시그모이드 함수를 사용하고 정량적이라면 선형 함수를 사용한다.

LSTM 순환 신경망을 이루는 셀의 내부는 Fig. 5[19]에 나타낸 것과 같다. 셀 스테이트 C_t는 과거 정보를 전달하는 역할을 하는 순환 신경망에서 가장 중요한 요소로 이전 셀에서 전달받은 셀 스테이트 C_{t-1}과 게이트에 의해 반영된 현재 셀의 정보에 의해서 업데이트된다. 게이트는 앞에서 언급했듯이 순환 신경망의 장기 의존성 문제를 해결해주며 입력 게이트 i_t, 망각 게이트 f_t, 출력 게이트 o_t로 이루어져 있고 현재 셀의 정보를 얼마나 반영할지 결정짓는 중요한 요소

이다. 현재 셀의 게이트, 셀 스테이트, 출력 연산은 이전 셀의 출력, 현재 입력을 연결한 $[h_{t-1},x_t]$ 와 가중치, 편향치를 선형 결합하고 게이트와 활성함수를 통과시키는 형태로 이루어지고 아래 Eqs. (6)-(11)[17,18]으로 나타낼 수 있다. 활성함수는 Eqs. (12)와 (13)으로 나타낸 하이퍼탄젠트 혹은 소프트사인[20]함수를 사용하며 본 연구에서는 소프트사인을 사용하였다.

$$i_t = \sigma(W_i[h_{t-1},x_t]+b_i), \tag{6}$$

$$f_t = \sigma(W_f[h_{t-1},x_t]+b_f), \tag{7}$$

$$o_t = \sigma(W_o[h_{t-1},x_t]+b_o), \tag{8}$$

$$\widetilde{C_t} = activation(W_c[h_{t-1},x_t]+b_c), \tag{9}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C_t}, \tag{10}$$

$$h_t = o_t \times activation(C_t). \tag{11}$$

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{12}$$

$$softsign(x) = \frac{x}{1+|x|}. \tag{13}$$

신경망의 학습은 다음과 같이 수행되었다. 본 연구에서는 피실험자로부터 취득한 4채널 초음파 도플러 신호의 각 채널별로 추출한 4×16 포인트 특징변수 프레임을 연속된 5개 프레임으로 연결한 5×64 형태의 행렬을 신경망의 입력으로, 그리고 입력 마지막 특징변수 프레임에 대응되는 음성 신호 프레임으로부터 추출한 128 포인트 푸리에 변환 계수를 신경망의 기준값으로 사용하였다. 전체 데이터의 70 %를 학습 데이터, 30 %를 테스트 데이터로 나누어 학습 데이터를 신경망 학습에 사용하였다. 신경망의 학습은 기준값과 신경망의 추정값에 대한 평균 제곱오차(Mean Squared Error, MSE)로 주어지는 손실 함수 L을 최소화되도록 진행되며 이에 따라 신경망 내 모든 변수들은 최적화되어졌다. N_{tr}개의 학습 데이터에서 k-번째 샘플을 신경망의 입력으로 사용할 때, 이에 대응되는 기준값과 추정값의 n-번째 값을 각각 $|y_k[n]|$, $\widehat{y_k}[n]$라 한다면 손실 함수 L은 다음 Eq. (14)와 같이 나타낼 수 있다.

$$L = \frac{1}{128N}\sum_{k}^{N_{tr}}\sum_{n=1}^{128}\left[\ |y_k[n]| - \widehat{y_k}[n]\ \right]^2. \tag{14}$$

III. 실험 및 결과

3.1 데이터 취득

실험을 위한 데이터를 취득하기 위해 초음파 방사를 위한 중심 주파수 40 kHz를 갖는 소형 초음파 트랜스듀서(AW8TR40, Audiowell, China, 음압레벨 115 dB), 초음파 도플러 신호 수신을 위한 광대역 특성의 초소형 MEMS 센서(SPM0404UD5, Knowles Acoustic, Japan, 수신감도 -47 dB, 10 kHz ~ 65 kHz), 음성 취득을 위한 가청주파수 대역의 마이크로폰(AKG880, AKG, Austria), 그리고 데이터 취득 위치가 변동되지 않기 위해 레이저 포인터를 미간에 위치하도록 사용했으며 센서와 장비는 Fig. 6과 같이 구성되었다. 40 kHz 초음파 신호는 함수발생기(33250A, agilent, USA)를 사용하여 발생되었고 마이크로폰과 4채널 초음파 센서에서 취득한 음성신호, 초음파 도플러 신호들은 다채널 오디오인터페이스(Fireface 800, RME, Germany)를 사용해 샘플링 주파수 192 kHz, 16비트 양자화로 채널 이득이 동일하도록 설정하여 동시에 디지털 값으로 변환되었다. 변환된 초음파 도플러 신호와 그에 동반하는 음성신호는 Fig. 7와 같이 나타난다. 실험 데이터는 문장 읽기를 정상적으로 수행할 수 있는

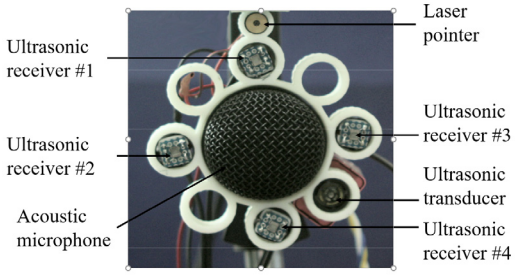

Fig. 6. Configuration of the acoustic Doppler microphone.

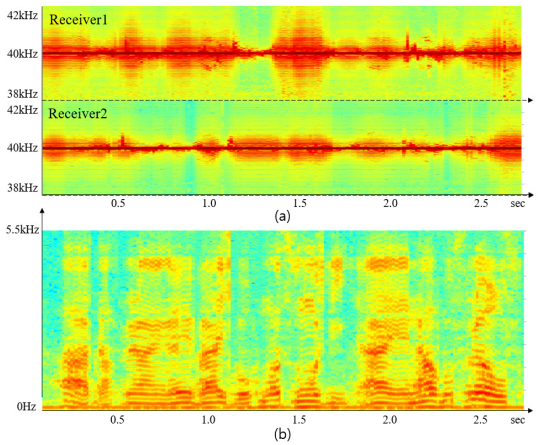

Fig. 7. An example of the spectrograms of (a) received ultrasonic signal, (b) corresponding speech signal.

건강한 피실험자 1명으로부터 취득되었다. 피실험자는 1063개의 문장 읽기를 수행하였으며 이로부터 취득된 음성신호에 101개 잡음환경을 더해 만들어진 13411개의 데이터와 초음파 도플러 신호로부터 특징변수를 추출하여 신경망의 입력 및 기준값으로 사용하였다.

LSTM 순환 신경망은 우분투(Ubuntu) 14.04, OS 환경에서 파이썬(Python) 3.6, 텐서플로우(Tensorflow) 1.8.0를 이용하고 Github의 Reference [21]를 참고하여 만든 소스코드에 의해 학습되었다. 학습에 사용된 그래픽 처리 장치(Graphics Processing Unit, GPU)로는 엔비디아 타이탄 XP(Titan XP, NVIDIA, U.S.A)를 사용하였다.

3.2 성능 평가

성능 평가 기준을 위하여 모든 데이터의 기준값과 추정값을 정규화하였다. 학습 데이터 기준값의 n-번째 포인트 평균과 표준편차를 각각 $\overline{y[n]}$, $\sigma[n]$라 할 때 정규화된 k-번째 기준값과 추정값의 n-번째 포인트 $|z_k[n]|$, $\hat{z}_k[n]$는 Eqs. (15)와 (16)과 같이 나타낼 수 있다. 실험의 객관적인 척도로는 Eq. (17)로 주어지는 정규화된 기준값과 추정값의 평균 제곱근 오차(Root Mean Squared Error, RMSE)가 사용되었다.

$$|z_k[n]| = \frac{|y_k[n]| - \overline{y[n]}}{\sigma[n]}, \tag{15}$$

$$\hat{z}_k[n] = \frac{\hat{y}_k[n] - \overline{y[n]}}{\sigma[n]}, \tag{16}$$

$$RMSE = \sqrt{\frac{1}{128N}\sum_{k}^{N}\sum_{n=1}^{128}\left[\ |z_k[n]| - \hat{z}_k[n]\ \right]^2}. \tag{17}$$

실험은 학습된 신경망에 학습, 테스트 데이터를 입력으로 주었을 때, 얻어지는 RMSE를 비교하며 진행되었다. 동일한 환경에서 LSTM과 MLP를 비교하기 위해 모든 실험은 101 epoch 학습했고 5개 시간 프레임을 입력으로 사용하는 LSTM과 동일한 입력 데이터 형태를 가지도록 연속된 5개 샘플을 연결하여 MLP의 입력 데이터로 사용하였다. 실험은 LSTM의 은닉 노드 수 변화에 따른 결과 비교를 통하여 LSTM에 사용할 은닉 노드 수를 결정하고 MLP, LSTM 각각 신경망의 레이어 수 변화에 따른 결과 비교, 신경망의 입력으로 사용되는 초음파 도플러 신호의 채널 수 변화에 따른 결과 비교로 진행되었다. LSTM과 MLP 모두 1920개 노드, 3 레이어, 초음파 도플러 신호 4개 채널을 사용한 실험에서 학습 데이터, 테스트 데이터를 각각 사용했을 때의 RMSE는 LSTM의 경우 각각 0.5170, 0.5810으로 나타났고, MLP의 경우 각각 0.5876, 0.7491로 나타났다. 모든 실험에서 평균적으로 LSTM에서는 약 0.06 그리고 MLP에서는 약 0.18의 RMSE 차이로 학습 데이터를 사용할 경우가 더 낮은 RMSE를 나타냈다. Fig. 8은 3 레이어 LSTM에 대해서 입력으로 4채널 초음파 도플러 신호를 사용했을 때, 은닉 노드 수를 640, 960, 1280, 1920개로 변화시킴에 따른 RMSE를 나타낸다. 은닉 노드 수를 640개, 1920개로 했을 때, RMSE는 각각 0.7304, 0.5810으로 약 0.15 차이를 나타낸다. 결과로부터 은닉 노드 수를 증가시킬수록 성능이 향상되는 것을 알 수 있다. 하지만 증가시키는 노드 수에 비례해서 학습 시간 또한 한 세대(Epoch)당 약 250 s에서 1700 s로 증가하기 때문에 학습 시간과 성능을 상황에 맞게 고려하여 은닉 노드 수를 결정해야 할 것으로 판단된다.

MLP, LSTM 두 신경망의 레이어 수 변화에 따른 RMSE는 Fig. 9에서 나타난다. 실험을 위하여 MLP, LSTM 모두 각 레이어의 노드 수를 1920개로 설정하고 입력으로 4채널 초음파 도플러 신호를 사용했다.

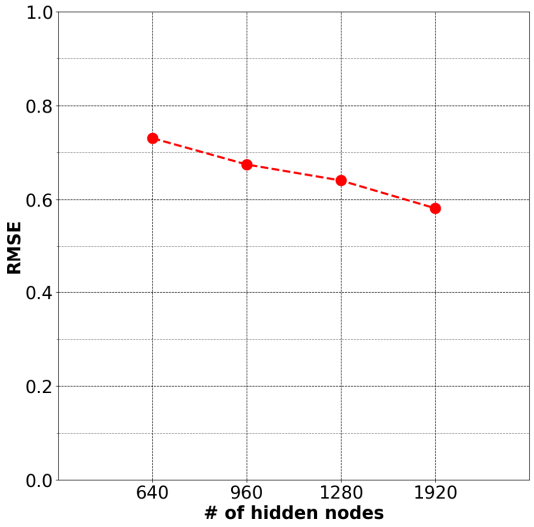

Fig. 8. RMSE of LSTM according to the number of hidden nodes.

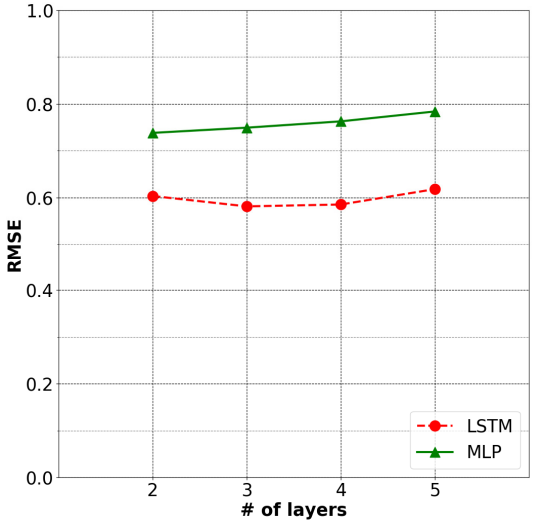

Fig. 9. RMSE of LSTM and MLP according to the number of layers.

실험을 진행한 결과 MLP의 경우 레이어 수가 2, 3, 4, 5개일 때, RMSE가 각각 0.7380, 0.7491, 0.7626, 0.7837로 2 레이어와 5 레이어를 사용할 때 RMSE가 최대, 최소로 약 0.046 차이나고 이에 따라 MLP의 레이어 수가 증가할수록 성능이 악화하는 결과를 나타냈다. LSTM의 경우 레이어 수가 2, 3, 4, 5개일 때, RMSE가 각각 0.6031, 0.5810, 0.5850, 0.6178로 MLP와는 달리 레이어 수 증가에 따른 성능 악화가 나타나지 않았고, 3 레이어를 사용했을 때 RMSE가 0.5810으로 가장 뛰어난 성능을 나타냈다. 그리고 LSTM이 MLP에

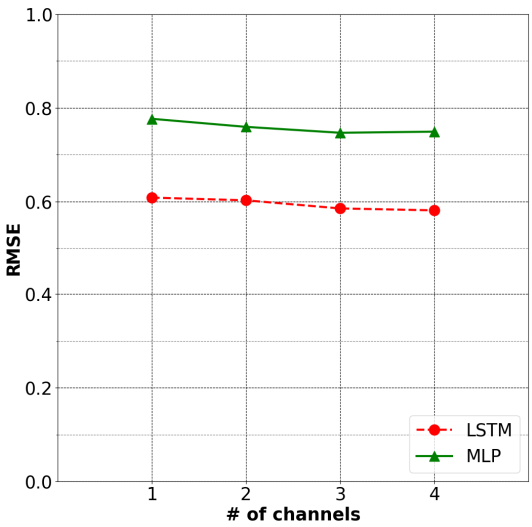

Fig. 10. RMSE of LSTM and MLP according to the number of ultrasonic Doppler signal channels.

비해 평균적으로 0.1616 더 낮은 RMSE를 가지는 것으로 성능 우세를 나타냈다.

Fig. 10은 두 신경망의 입력 데이터로 사용되는 4 채널 초음파 도플러 신호의 채널 수에 따른 RMSE를 나타낸다. 각 채널마다 길이 16의 특징변수를 가지고 이에 따라 입력 데이터 길이는 최소 16, 최대 64가 될 수 있다. MLP, LSTM 모두 각 레이어의 노드 수를 1920개 3 레이어로 설정해 사용했다. 실험을 통하여 LSTM의 경우 채널을 1, 2, 3, 4개 사용할 때 RMSE가 각각 0.6082, 0.6024, 0.5851, 0.5810로 나타나고 MLP의 경우 RMSE가 각각 0.7765, 0.7593, 0.7467, 0.7491로 나타났다. 이에 따라 LSTM과 MLP 두 신경망 모두 채널 수가 증가할수록 성능이 향상되는 것을 나타내고 LSTM이 MLP에 비해 평균적으로 0.1609 더 낮은 RMSE를 가지는 것으로 성능 우세를 나타냈다. 위 실험들을 통하여 LSTM과 MLP 두 신경망이 가장 우수한 성능을 나타낼 수 있는 초음파 도플러 신호 채널 수, 은닉 노드 수, 레이어 수를 확인할 수 있었다. 그리고 이를 이용해 학습시킨 두 신경망은 RMSE가 각각 0.5810, 0.7380로 나타나 0.1570 더 낮은 RMSE를 가지는 LSTM의 성능이 MLP에 비해 객관적으로 우수하다는 것을 확인할 수 있었다. 객관적 척도를 이용한 성능평가에 따른 두 신경망의 성능차이가 추정결과에서 어떠한 형태로 나타나는지 확인하기 위해 두 신경망의 추정결과를 비교하였다. Fig. 11은 두 신경

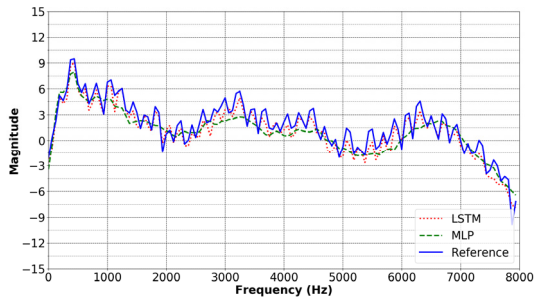

Fig. 11. Comparison of MLP and LSTM feature variables estimation.

망을 이용해 음성 특징변수를 추정한 결과를 나타낸다.

결과로부터 LSTM과 MLP 두 신경망 모두 전체적으로 기준값에 유사하게 추정할 수 있음을 확인할 수 있다. 그러나 LSTM으로 추정하는 경우 유성음 구간에서 나타나는 특징인 고조파 구조가 뚜렷하게 나타나는 반면, MLP로 추정하는 경우 스펙트럼 포락선은 유사하지만 고조파 구조가 소실되어 나타났다. 따라서 전체적인 추세를 나타내는 스펙트럼 포락선보다는 고조파 구조를 나타내는 세밀한 부분에 대한 추정능력에서 두 신경망의 성능차이가 나타나는 것을 확인할 수 있다.

IV. 결 론

LSTM 순환 신경망을 이용해 음성신호 없이 4채널 초음파 도플러 신호로부터 추정한 신호로부터 충분히 음성합성이 가능함을 보였다. 기존 MLP를 이용한 방법과 비교해 LSTM 순환 신경망은 음성신호와 같은 시계열 데이터의 특성을 고려한 구조를 지녔기 때문에 LSTM 순환 신경망을 이용한 방법이 더욱 우수한 성능을 가진다는 것을 보여주었다. 본 연구에서는 문장으로부터 취득한 데이터를 이용하였지만 실질적으로 음성합성, 음성인식 기술 적용을 검토하기 위해서 음소, 음절 단위의 데이터들을 이용한 연구가 필요하다고 판단된다.

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. K. S. Lee, “Prediction of acoustic feature parameters using myoelectric signals,” IEEE Trans. On Biomed. Eng. **51**, 1587-1595 (2010).
3. T. Toda and K. Shikano, “NAM-to-Speech conversion with Gaussian Mixture Models,” Proc. Interspeech, 1957-1960 (2005).
4. S. Li, J. Q. Wang, M. Niu, T. Liu, and X. J. Jing, “The enhancement of millimeter wave conduct speech based on perceptual weighting,” Progress in Electromagnetics Research B, **9**, 199-214 (2008).
5. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Comm. **54**, 134-146 (2012).
6. K. Kalgaonkar and B. Raj, “An acoustic Doppler-based front end for hands free spoken user interaces,” Proc. SLT, 158-161 (2006).
7. K. Kalgaonkar and B. Raj, “Acoustic Doppler sonar for gait recognition,” Proc. 2007 IEEE Conf. Advanced Video and Signal Based Surveillance, 27-32 (2007).
8. K. Kalgaonkar and B. Raj, “One-handed gesture recognition using ultrasonic Doppler sonar,” Proc. ICASSP, 1889-1892 (2009).
9. S. Srinivasan, B. Raj, and T. Ezzat, “Ultrasonic sensing for robust speech recognition,” Proc. ICASSP, 5102-5105 (2010).
10. K. Livescu, B. Zhu, and J. Glass, “On the phonetic information in ultrasonic microphone signals,” Proc. ICASSP, 4621-4624 (2009).
11. A. R. Toth, B. Raj, K. Kalgaonkar, and T. Ezzat, “Synthesizing speech from Doppler signals,” Proc. ICASSP, 4638-4641 (2010).
12. K. S. Lee, “Speech synthesis using acoustic Doppler signal” (in Korean), J. Acoust. Soc. Kr. 35, 134-142 (2016).
13. K. S. Lee, “Automatic speech recognition using acoustic doppler signal” (in Korean), J. Acoust. Soc. Kr. **35**, 74-82 (2016).
14. K. S. Lee, “Speech Enhancement using ultrasonic doppler sonar,” Speech Comm. **110**, 21-32 (2019).
15. F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (Spartan Books, Washington DC, 1961), pp. 3-585.
16. D.E. Rumelhart, G.E. Hilton, and R.J. Williams, “Learning internal representations by error propagation,” *in parallel distributed processing: Explorations in the microstructure of cognitio* (MIT press, Cambridge, 1986), pp. 318-362.
17. S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation. **9**, 1735-1780 (1997).

18. F. Gers, N. Schraudolph, and J. Schmidhuber, “Learning Precise Timing with LSTM Recurrent Networks,” Journal of Machine Learning Research. **3**, 115-143 (2002).
19. *Understanding LSTM Networks*, http://colah.github.io/posts/2015-08-Under standing- LSTMs/, 2019.
20. J. Turian, J. Bergstra, and Y. Bengio, “Quadratic features and deep architectures for chunking,” Proc. NAACL HLT 2009, 245-248 (2009).
21. *Ptb_Word_lm.py*, https://github.com/tensorflow/models/blob/master/tutorials/rnn/ptb/ptb_word_ lm.py

저자 약력

▸ 주 형 길 (Hyeong-Kil Joo)

2018년 2월: 건국대학교 전자공학과 학사
2018년 3월 ~ 현재: 건국대학교 전자공학과 석사과정

▸ 이 기 승 (Ki-Seung Lee)

1991년 2월: 연세대학교 전자공학과 학사
1993년 2월: 연세대학교 전자공학과 석사
1997년 2월: 연세대학교 전자공학과 박사
2000년 9월: AT&TLabs-Research, Senior technical staff member
2001년 8월: 삼성전자(주)종합기술원
2001년 9월 ~ 현재: 건국대학교 전자공학과 교수