Journal of The Institute of Electronics and Information Engineers Vol.54, NO.5, May 2017 https://doi.org/10.5573/ieie.2017.54.5.85
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2017-54-5-10

음소기반의 순환 신경망 음성 검출기를 이용한 음성 향상

(Speech Enhancement using RNN Phoneme based VAD)

이 강*, 강 상 익*, 권 장 우*, 이 상 민**

(Kang Lee, Sang-Ick Kang, Jang-woo Kwon, and Samgmin Lee©)

요 약

본 논문에서는 향상된 연산 능력을 가진 하드웨어와 알고리즘의 혼합을 통하여 음성 향상을 위한 정확한 음성 검출기 구현을 목적으로 하였다. 음성은 음소의 나열로 구성되어있으며 음성 모델을 세우는데 적합한 방법은 이전의 정보를 이용하는 순환 신경망 (recurrent neural network, RNN)을 사용하는 것이다. 실제 존재하는 모든 잡음에 대하여 학습한 모델을 제시하는 것은 사실상 불가능 하므로 이를 극복하고자 음소기반 학습을 진행하였다. 학습의 결과로 세워진 모델을 기반으로 새로운 음성 신호에서 음성을 검출하고 그 결과를 이용하여 음성 향상을 진행하였다. 순환 신경망과 음소기반 학습은 프레임 별 높은 상관성을 가진 음성 신호에서 좋은 성능을 얻을 수 있었으며 음성 검출기의 성능을 검증하기 위하여 라벨 데이터와 음성 검출 결과를 비교하고 다양한 잡음 환경에서 객관적 음질 평가를 진행하여 기존의 음성 향상 알고리즘과 비교하였다.

Abstract

In this papers, we apply high performance hardware and machine learning algorithm to build an advanced VAD algorithm for speech enhancement. Since speech is made of series of phoneme, using recurrent neural network (RNN) which consider previous data is proper method to build a speech model. It is impossible to study every noise in real world. So our algorithm is builded by phoneme based study. we detect voice present frames in noisy speech signal and make enhancement of the speech signal. Phoneme based RNN model shows advanced performance in speech signal which has high correlation among each frames. To verify the performance of proposed algorithm, we compare VAD result with label data and speech enhancement result in various noise environments with previous speech enhancement algorithm

Keywords : RNN, GMM, Phoneme, VAD, MMSE

Ⅰ. 서 론

음성과 비음성 구간을 구별하는 전통적인 문제를 해결하는 음성 검출기 (voice activity dection, VAD)는 음성 코덱 (codec), 음성 인식 (speech recognition), 음성 향상 (speech enhancement), 반향 제거 (echo cancellation) 등 많은 응용분야에서 사용된다. 최근에는 서버 기반의 음성 인식 시스템의 발전으로 접근성이 향상된 음성 인식 분야는 음성 검출기의 커다란 적용 분야이다. 또한 화자 인식 (speaker verification)과 감정 인식 (emotional speech recognition) 등 새로운 응용 분야에서 음성 검출기가 적용되기도 한다. 대부분의 시스템은 장소 제약 없이 사용되어지기 때문에 이 때 입력되는 음성은 다양한 잡음에 영향을 받아 왜곡되어지며 시스템은 왜곡되어진 신호에 민감하여 성능이 큰 폭으로 감소되기 때문에 이를 해결하기 위해 음성 검출기 기반의 음성 향상 알고리즘을 필요로 한다[1].

음성 검출기 기반의 음성 향상 알고리즘은 음성 검출기를 이용하여 음성 부재 구간을 얻고 비음성 구간에서 잡음의 평균 전력을 구할 수 있다. 잡음의 평균 전력을 가지고 오염된 음성 신호의 스펙트럼 크기에 잡음 제거 이득을 곱하여 깨끗한 음성을 추정한다.

* 정회원, ** 평생회원 인하대학교 전자공학과, 컴퓨터공학과
(Dept. of Electronic, Computer Information Engineering, Inha University)
© Corresponding Author(E-mail : sanglee@inha.ac.kr)
※ 이 논문은 2010년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2010-0020163)
※ 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.NRF-2016R1A2B4015370)
Received ; December 28, 2016 Revised ; February 6, 2017
Accepted ; April 12, 2017

음성 향상 알고리즘에 적용하는 음성 검출기는 많은 연구가 진행되었으며, 신경망(neural network, NN)을 이용하여 시간-주파수 분류기에 적용되는 마스크를 추정하는 서포트 벡터 머신(support vector machine, SVM)을 학습시키는 방법[2]과 비교사 학습(unsupervised learning)의 한 종류인 기댓값 최대화 알고리즘(expectation-maximization algorithm, EM algorithm)을 이용하여 음성 모델의 파라미터들을 학습하는 방법[3] 등이 존재하였다. 하지만 특정한 모델이나 가정을 이용하지 않았기에 학습되지 않은 잡음 환경에 대해서 낮은 성능을 나타내었으며 존재하는 모든 잡음 환경에 대하여 학습을 진행하는 것은 불가능하다는 점과 학습한 음성 모델이 음성을 이루는 음소의 특징에 대하여 고려하지 않았다는 문제가 존재한다.

본 논문에서는 음소별로 표시된 데이터베이스를 이용하여 순환 신경망(recurrent neural network, RNN)을 학습시키며 음성의 존재 확률(speech presence probability, SPP)를 계산한다. 본 알고리즘은 이전 단계에서의 결과를 고려하는 순환 신경망을 도입하였기 때문에 연속적인 특성을 가진 음성 모델링에 이점을 갖는다. 생성 모델의 한 종류인 순환 신경망을 사용하여 나타나는 단점을 판별 모델의 한 종류인 가우시안 혼합 모델 (Gaussian mixture model, GMM)을 이용하여 보완하였다. 향상된 음성 검출기를 이용하여 MMSE (minimum mean-square error) 방법으로 오염된 음성 신호의 향상을 진행하였으며[4], 향상된 결과를 객관적 음질평가 방법인 PESQ(perceptual evaluation of speech quality)를 이용하여 검증하였다[5].

Ⅱ. 음성 검출기를 이용한 음성 향상

오염된 음성 신호는 깨끗한 음성과 잡음이 동시에 존재한다고 간주 할 수 있다. 간단한 잡음 모델은 다음과 같이 나타낼 수 있다.

$$z(t) = x(t) + y(t) \tag{1}$$

여기서, $z(t)$는 오염된 음성신호이며, $x(t)$는 깨끗한 음성신호, $y(t)$는 가산 잡음을 나타낸다. 이 신호들은 모두 시간 t에 따라 변하게 된다. 음성 검출기는 주파수 축에서 동작하며 국소 푸리에 변환 (short-time Fourier transform, STFT)를 이용하여 다음과 같이 변환 시킬 수 있다.

$$Z(n,k) = X(n,k) + Y(n,k) \tag{2}$$

여기서 $Z(n,k)$, $X(n,k)$, $Y(n,k)$는 $z(t)$, $x(t)$, $y(t)$를 프레임 길이 L 샘플만큼 잘라 오버랩하여 스펙트럼으로 변환한 결과이며, n은 프레임의 인덱스 (frame index)이며 k는 frequency bin이다. 일반적으로 프레임 인덱스 n은 편의를 위하여 생략한다.

이를 이용하여 음성 존재 구간과 음성 부재 구간에 대한 가정을 다음과 같이 나타낼 수 있다.

$$\begin{aligned} H_0 &: speech\,absent : Z(k) = Y(k) \\ H_1 &: speech\,present : Z(k) = X(k) + Y(k). \end{aligned} \tag{3}$$

시간-주파수 영역에서의 음성신호 모델을 나타내는 식 (2)에서 오염된 음성신호 스펙트럼 $Z(k)$의 극좌표 형태 표현은 다음과 같다.

$$Z(k) = |Z(k)|e^{j\phi_z(k)} \tag{4}$$

여기서, $\phi_z(k)$는 오염된 음성 신호의 위상을 나타내고 $|Z(k)|$는 스펙트럼의 크기이다. 일반적으로 사람의 귀는 스펙트럼의 크기의 변화에 민감하고 위상 변화에 대해서는 둔감하다. 따라서 오염된 음성 신호의 스펙트럼 크기에 잡음제거 이득을 곱하여 깨끗한 음성 신호의 스펙트럼 크기를 추정하고, 위상 값은 오염된 음성 신호의 위상 값을 수정 없이 사용한다.

$$\hat{X}(k) = |G(k)Z(k)|e^{j\phi_y(k)} \tag{5}$$

$\hat{X}(k)$는 추정된 깨끗한 음성 신호의 주파수 스펙트럼을 나타내고, $G(k)$는 k 번째 frequency bin에서의 잡음제거 이득을 나타낸다.

최소 평균 제곱 오차 (minimum mean square error, MMSE) 기반으로 오염된 음성신호에서 추정된 깨끗한 신호를 구하는 기댓값은 $\hat{x} = E(X|Z = z)$로 나타낼 수 있다.

Ⅲ. 제안된 음소기반 순환 신경망 음성 검출기를 이용하는 음성향상 알고리즘

음성 검출기는 정확한 음성존재확률을 계산하는 것을 목적으로 한다. 음성존재확률은 음소의 사후확률 (posterior probability) $p_i = p(I = i|Z = z)$에 의하여 계산되어진다. 본 논문에서는 판별모델인 순환 신경망을 이용하여 교사학습 (supervised learning)을 진행하여 음소를 검출해내어 음성존재확률을 계산한다. 순환

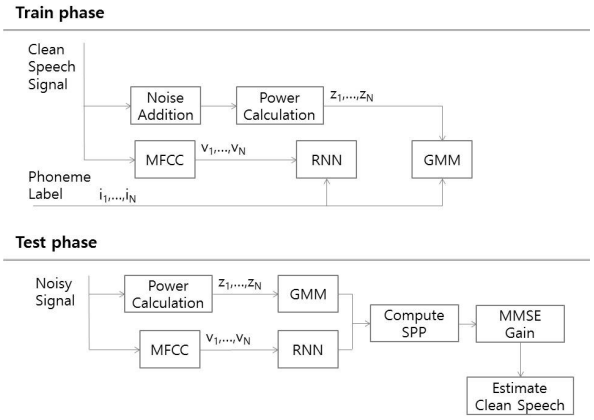

그림 1. 제안된 알고리즘의 전체 블록도
Fig. 1. Block diagram of the proposed method.

신경망의 은닉층 (hidden layer)는 이전 시간의 값과 현재 시간의 입력 값에 의하여 계산되며 다음과 같이 나타낼 수 있다.

$$s(t) = f(U_{x(t)} - W_{s(t-1)}) \tag{6}$$

여기서 f는 비선형 판별함수를 나타내는데 ReLU (Rectified Linear Unit)를 사용하였다. 출력층은 클래스의 수에 대한 확률 벡터로 나오게 되는데 이는 다음과 같이 나타낼 수 있다.

$$y(t) = softmax(V_{s(t)}) \tag{7}$$

여기서 softmax 함수는 입력이 어떤 클래스에 속한다는 증거를 더한 다음 해당 증거를 확률로 변환시켜 출력을 구해준다.

음소기반의 순환 신경망을 학습하기 위하여 깨끗한 음성과 음소가 라벨링 된 데이터를 이용한다. 라벨링 된 음소 데이터는 TIMIT 데이터베이스를 이용하였다[6]. TIMIT의 음소 데이터는 프레임별 우세한 음소 성분을 61개의 클래스로 나타내고 있는데, 이를 학습에 사용하기 위하여 음소 클래스 중 유사한 클래스를 묶어 39개의 클래스로 줄여 학습에 사용하였다[7]. 음성은 13차 MFCC (mel-frequency cepstral coefficients)를 이용하여 특징을 추출하고 9개의 프레임(이전 4프레임, 현재 프레임, 이후 4프레임)의 MFCC 벡터를 합쳐 이용한다. 음소기반의 순환신경망 음성 검출기는 다음과 같이 음성 존재확률을 구한다.

$$p_i^{RNN} = p(I = i|v; RNN) \tag{8}$$

여기서 v는 9개의 프레임에서의 MFCC 특징벡터이며, 순환 신경망 구조를 이용하여 실시간 환경에서의 속적인 데이터에 대한 모델링이 가능하다.

순환 신경망을 보완하기 위하여 적용한 GMM은 하나의 단일 가우시안이 하나의 음소 클래스와 대응되므로 음소의 클래스 수인 39개의 가우시안의 혼합으로 이루어져 있으며, 각각의 frequency bin에서의 평균과 분산을 계산한다. 각 frequency bin에서의 음소 라벨 i에 대한 평균과 분산은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \mu_{i,k} &= \frac{1}{N_i} x_{i,k}(n) \\ \sigma_{i,k}^2 &= \frac{1}{N_i - 1} \sum_{n=1}^{N_i} (x_{i,k}(n) - \mu_{i,k})^2 \end{aligned} \tag{9}$$

여기서 $x_{i,k}(n)$은 k번째 frequency bin에서의 i번째 음소에 대한 n번째 로그 스펙트럴 벡터 (log spectral vector)이다. 혼합 계수 c_i는 학습 데이터에서 각 음소의 상대적 빈도수로 나타낼 수 있다.

$$c_i = \frac{N_i}{\sum_{n=1}^{m} N_n} \tag{10}$$

학습된 GMM과 RNN의 혼합은 MMSE 방법과 음성 존재확률을 동시에 이용한다. MMSE 방법에 사용되는 깨끗한 음성의 추정값은 다음의 과정을 통하여 정의된다.

$$\hat{x} = \sum_{i=1}^{m} p(I = i|Z = z) E(X|Z = z, I = i) \tag{11}$$

추정된 잡음을 주파수 성분을 고려하면 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \widehat{x_{i,k}} &= E(X_k|Z_k = z_k, I = i) \\ &= \rho_{i,k} z_k + (1 - \rho_{i,k}) E(X_k|X_k < z_k, I = i). \end{aligned} \tag{12}$$

식 (8)과 식 (11)을 이용하여 음성 존재 확률인 ρ_k를 다음과 같이 구할 수 있다.

$$\rho_k = \sum_{i=1}^{m} p_i^{RNN} \rho_{i,k} \tag{13}$$

GMM 음성 검출기에 적용을 위해 잡음의 평균과 분산 값의 평균이 필요하다. 잡음의 평균과 분산의 평균을 추정하기 위하여 음성이 시작되는 부분(보통 0.25초) 내에 잡음 추정이 중단되고 이때까지의 정보를 가지고 다음과 같이 가우시안 평균과 분산의 초기값을 구한다.

$$\mu_{Y,k} = \frac{1}{N_Y}\sum_{n=1}^{N_Y} y_k(n)$$
$$\sigma^2_{Y,k} = \frac{1}{N_Y - 1}\sum_{n=1}^{N_Y}(y_k(n)-\mu_{Y,k})^2 \quad (14)$$

N_Y는 잡음만 존재 하였을 때의 잡음의 종류의 수이며, $y_k(n)$은 k번째 bin의 n번째 잡음 벡터이다.

잡음만 존재하는 잡음의 평균과 분산의 갱신은 식 (13)에서 구한 음성 존재 확률을 이용하여 다음과 같이 구할 수 있다.

$$\mu^{new}_{Y,k} = \rho_k \mu^{old}_{Y,k} + (1-\rho_k)(\alpha z_k + (1-\alpha)\mu^{old}_{Y,k})$$
$$\sigma^{new}_{Y,k} = \rho_k \sigma^{old}_{Y,k} + (1-\rho_k)\left(\alpha\sqrt{(z_k-\mu^{new}_{Y,k})^2} + (1-\alpha)\alpha^{old}_{Y,k}\right) \quad (15)$$

$\mu^{new}_{Y,k}$와 $\sigma^{new}_{Y,k}$는 갱신된 값이며 $0<\alpha<1$은 스무딩 파라미터 (smoothing parameter)이다. 식 (15)를 통한 갱신은 음성이 존재하는 경우에도 적용가능하며, 비정상적인 잡음 환경에서 특별히 좋은 성능을 갖는다.

Ⅲ. 실험 및 결과

본 장에서는 제안한 VAD 알고리즘을 이용하여 오염된 음성신호의 향상을 다룬다. TIMIT 데이터와 NOISEX-92의 잡음 데이터를 이용하여 음성신호를 오염시켰으며, Babble, Volvo, White, Factory 잡음을 5dB, 10dB, 15dB SNR 레벨로 깨끗한 음성신호에 적용시켰다[8]. 학습은 전체 데이터의 73%를 이용하였으며 나머지 27%를 이용하여 테스트를 진행하였다. 기존에 널리 사용되는 IMCRA (improved minima controlled recursive averaging) 잡음 추정기 알고리즘의 음성 향상 결과와 비교하였으며[9], 향상된 신호의 객관적인 평가를 위하여 객관적 음질 평가 방법인 PESQ (perceptual evaluation of speech quality)를 이용하였다.

표 1은 Babble, Volvo, White, Factory 잡음에서의 오염된 음성신호, IMCRA를 이용하여 음성 향상된 신호, 본 논문에서 제안한 알고리즘을 이용하여 음성 향상된 신호의 PESQ 결과를 나타낸다. 결과를 살펴보면 다양한 잡음 환경에서 기존의 알고리즘과 본 논문에서 제안한 알고리즘 모두 오염된 신호보다 높은 점수를 획득하였다. 또한 다양한 잡음환경에서 기존의 음성 향상 알고리즘보다 본 논문에서 제안한 알고리즘의 향상 폭이 더 크게 나타났다. 전체 실험에서의 기존의 알고리즘의 평균은 11.6% 이었으며 제안한 알고리즘의 평균은 14.0%였다. 음질 평가 결과의 상승률은 기존 방법대비 약 20.6% 증가하였다.

표 1. 다양한 잡음환경에서 기존과 제안한 알고리즘에 대한 PESQ 수치

Table1. PESQ scores of the conventional method and the proposed method under noise environments.

Noise	SNR (dB)	Method		
		Noisy	IMCRA	Proposed
Babble	5	2.214	2.325	2.411
	10	2.447	2.717	2.794
	15	2.768	3.154	3.172
Volvo	5	2.381	2.611	2.683
	10	2.612	2.916	2.935
	15	2.819	3.214	3.238
White	5	2.114	2.542	2.558
	10	2.431	2.733	2.764
	15	2.683	2.904	2.928
Factory	5	1.887	2.116	2.317
	10	2.302	2.665	2.697
	15	2.652	3.019	3.071

Ⅳ. 결 론

본 논문에서는 음성 인식, 음성 향상 등 다양한 응용분야에서 사용되는 음성 검출기의 성능을 향상 시켰다. 음성 검출을 위하여 기계학습 방법을 사용하였으며 기계학습 중 생성 모델인 가우시안 혼합 모델과 판별 모델인 순환 신경망을 도입하여 서로의 단점을 보완할 수 있도록 구성하였다. 또한 음소를 이용하여 학습하였기 때문에 학습되지 않은 잡음 환경에 대한 강인함을 가지고 있다.

가우시안 혼합 모델과 순환 신경망 모델은 각각 프레임과 frequency bin에서 음성 존재확률을 계산한다. 구해진 음성 존재확률을 기반으로 잡음의 전력을 추정할 수 있게 되며 이득 제어 알고리즘을 이용하여 깨끗한 음성의 전력을 추정할 수 있다. 음성 검출기의 성능은 향상된 신호의 음질 평가를 진행하여 검증하였다.

객관적 음질 평가 방법을 사용하여 기존의 음성 향상 알고리즘과 비교하였으며 잡음 환경에 따라 성능의 차이는 있었지만 수행한 환경에서 기존 알고리즘 대비 향상된 결과를 보여주었다.

REFERENCES

[1] H. G. Hirsch, and D. Pearce. “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW). 2000.
[2] Y. Wang and D. Wang, “Towards scaling up classification-based speech separation,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 12, pp. 1381-1390, July 2013.
[3] D. Burshtein, and S. Gannot, “Speech enhancement using a mixture-maximum model,” IEEE transactions on speech and audio processing, vol. 10, no. 6 pp. 341-351, 2002.
[4] Loizou, Philipos C. “Speech enhancement: theory and practice.” CRC press, 2013.
[5] A. W. Rix, J. G. Beerends, M. P. Hollier, P. Hekstra “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs.” Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. Vol. 2. IEEE, 2001.
[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, and J. G. Fiscus, “TIMIT acoustic-phonetic continuous speech corpus,” Linguistic data consortium, Philadelphia vol. 33, 1993.
[7] C. Lopes, and F. Perdigao. “Phone recognition on the TIMIT database,” Speech Technologies/Book 1, pp. 285-302, 2011.
[8] A. Varga and H. J. Steeneken, “Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech communication, vol. 12, no. 3, pp. 247-251, 1993.
[9] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” IEEE Transactions on speech and audio processing vol. 11, no. 5, pp. 466-475, 2003.
[10] Y. S. Park, H. S. Ahn, and S. M. Lee, “Speech Enhancement Based on Teager Energy and Speech Absence Probability in Noisy Environments.” IEIE Journal-SP, vol. 49. no. 13, pp. 81-88, 2012.

저 자 소 개

이 강(정회원)
2015년 인하대학교 생명공학과, 전자공학과 학사 졸업.
2017년 인하대학교 전자공학과 석사 졸업.
<주관심분야: Digital Signal Processing, Machine Learning, Speech Enhancement>

강 상 익(정회원)
2007년 인하대학교 전자공학과 학사 졸업.
2009년 인하학교 전자공학과 석사 과정 졸업.
2009년～현재 인하대학교 전자공학과 박사과정
<주관심분야: Machine Learning, 음성검출기>

권 장 우(정회원)
1990년 인하대학교 전자공학과 학사 졸업.
1992년 인하대학교 전자공학과 석사 졸업.
1996년 인하대학교 전자공학과 박사 졸업.
2006년～현재 인하대학교 컴퓨터공학과 교수.
<주관심분야: Human-Computer Interaction, Signal Processing, Intelligent System>

이 상 민(평생회원)
1987년 인하대학교 전자공학과 학사 졸업.
1989년 인하대학교 전자공학과 석사 졸업.
2000년 인하대학교 전자공학과 박사 졸업.
2006년～현재 인하대학교 전자공학과 교수.
<주관심분야: Bio-Signal Processing, Psyco-Acoustic, Brain-Machine Interface>