김일성종합대학학보
정보과학
주체109(2020)년
제66권 제1호

부름말검출에 의한 목적발성자의 음성강조

리지은, 곽청일

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《오늘 세계는 경제의 지식화에로 전환되고있으며 우리앞에는 나라의 경제를 지식의 힘으로 장성하는 경제로 일신시켜야 할 시대적과업이 나서고있습니다.》

자동음성인식(ASR)체계에서 주위의 잡음을 약화시키고 주목하는 발성자(목적발성자)의 음성만을 강조하는 문제는 체계의 인식정확도에 큰 영향을 미친다.

특히 가정들에서 리용되는 지능고성기에서 목적발성자의 음성을 강조하는 문제는 매우 중요하게 제기된다. 지능고성기를 비롯한 음성인식체계에서 음성강조문제는 배경잡음과 배경음성환경(목적발성자가 아닌 다른 발성자의 발성, 텔레비죤, 록음기, 고성기에서 나오는 록음이 존재하는 환경)에서 목적발성자의 음성을 추출, 강조하여 음성인식처리단에 보내는 문제이다. 이 음성강조문제는 아직까지 원만히 해결되지 못하고있다.

목적발성자의 음성강조문제는 보통 맹목음성분리에 기초하고있다.

한통로방법으로서 비부값행렬인수분해(NMF)방법[1], 시간－주파수마스크방법, 다통로방법으로서 독립성분분석(ICA), 심층학습에 기초한 방법[3] 등의 맹목음성분리방법들이 연구되였지만 분리성능이 높지 못하다. 특히 이 방법들에서는 치환불확정성문제 즉 분리된 신호들중에서 어느것이 목적발성자의 음성인가를 알아내는 문제를 풀어야 한다.

이 치환불확정성문제를 해결하려면 발성자의 특징이나 신호의 세기에 대한 사전지식이 있어야 한다. 선행연구[2]에서는 사전에 목적발성자의 깨끗한 음성을 록음한 자료를 가지고있다는 가정하에서 목적발성자의 음성신호를 찾아내는 한가지 방법(speaker beam)을 제안하였다.

우리는 론문에서 치환불확정성문제를 피하고 부름말검출에 의한 발성자의 음성을 강조하는 한가지 방법을 연구하였다. 이 방법은 어떤 특정한 부름말을 사전에 정하여 그것의 특징벡토르를 보관하고있다가 실지환경에서 부름말을 발성하는 사람을 목적발성자로 보고 공간빔형성기술을 리용하여 그 이후에 발성하는 음성을 강조한다.

1. 음성강조체계

목적발성자는 먼저 부름말을 발성한 다음에 음성인식을 해야 할 기본지령문을 발성한다고 가정한다.

론문에서 제안한 음성강조를 위한 체계는 여러개의 마이크들로 이루어진 수감부배렬을 리용한다.

우리가 제기하는 체계는 크게 두가지 단계 즉 마스크추정방법에 의한 부름말구역검출단계와 빔형성에 의한 음성강조단계로 이루어져있다.

첫 단계인 마스크추정방법에 의한 부름말구역검출단계에서는 심층신경망에 기초한 마스크추정방법을 리용하여 혼합신호를 부름말과 그 나머지 배경음성신호로 분리한다.

두번째 단계인 빔형성에 의한 음성강조단계에서는 분리된 신호를 가지고 빔형성려파

기를 계산하여 목적발성자로부터 련속적으로 발성되는 음성을 강조한다.

제안된 목적발성자음성강조체계의 흐름도를 그림에 보여주었다.

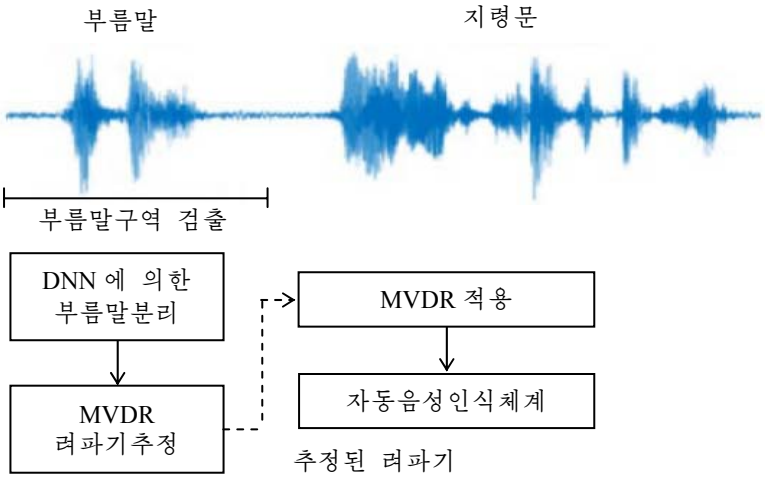

그림. 목적발성자음성강조체계의 흐름도

사전에 심층신경망기술을 리용하여 부름말과 비부름말(부름말이 아닌 배경음성 및 배경잡음)에 해당되는 시간－주파수마스크를 추정한다.

체계는 항시적으로 마이크배렬로 입력되는 혼합신호렬에서 부름말구역이 있는가를 검사한다.

일단 부름말구역이 검출되면 부름말마스크와 비부름말마스크를 원래의 혼합신호에 각각 적용하여 부름말과 그 나머지배경음성을 분리한다. 이 처리는 매 마이크통로에서 반복되며 분리된 다중통로신호들은 빔형성기를 계산하는데 리용된다. 빔형성기로서는 잘 알려진 최소분산무외곡응답(MVDR)빔형성기를 리용한다.

계산된 빔형성기를 리용하여 목적발성자의 음성을 강조하고 배경음성을 감쇠시킨다.

최소분산무외곡응답빔형성려파기는 부름말구역에서만 한번 계산되고 그 이후에 련이은 지령문구간에서는 갱신되지 않는다.

목적발성자의 음성이 강조된 음성신호는 자동음성인식체계의 입력으로 된다.

2. 마스크추정방법에 의한 부름말구역검출단계

부름말과 배경음성이 혼합된 신호가 입력으로 주어지면 심층신경망은 두가지 형태의 출력을 내보낸다. 한 형태는 부름말이며 다른 형태는 비부름말(배경음성)이다.

혼합된 신호에서 특징벡토르들을 추출한다. 표본화주파수는 16kHz, 한 프레임의 시간은 32ms, 프레임밀기시간은 16ms이며 하밍창문함수를 리용한다.

특징량으로서 스펙트로그람을 리용하며 매 프레임에서 256차원특징량을 추출한다.

문맥결합을 위하여 이웃한 20개의 문맥프레임(왼쪽과 오른쪽에서 각각 10개의 프레임)을 포함하며 결과적으로 5 376차원벡토르가 리용된다. 이 특징량들을 표준화한 다음 심층신경망의 입구에 넣는다.

심층신경망은 3개의 완전련결된 은닉층으로 구성하며 매 은닉층에는 1 024개의 마디점들이 있다.

출력층은 2개의 출력형태에 대하여 각각 256개의 출력마디들을 가지고있다.

은닉층에 리용된 활성화함수는 정규선형함수(ReLU)이며 출력층에서는 시그모이드함수를 써서 심층신경망의 출력값을 0부터 1까지의 범위로 제한한다.

심층신경망의 파라메터들은 2개의 출력형태들과 주어진 기준값사이의 오차를 최적화하면서 학습된다.

오차함수는 교차엔트로피함수를 리용하여 확률적그라디엔트하강법(SGD)으로 학습한다. 미니배치의 크기는 128로 설정한다. 신경망의 생략률은 입력층에 대하여 0.2의 값을 리용한다.

부름말과 배경음성은 같은 방의 조용한 환경에서 따로따로 록음하였다. 록음한 부름말음성은 35명의 발성자로부터 1 660개, 배경음성수는 25명으로부터 얻은 1 400개이다. 부름말음성과 배경음성을 여러가지로 결합시켜 도합 116 200개의 혼합음성을 만들었으며 이것을 DNN에 기초한 마스크추정을 위한 학습자료기지로 리용하였다.

혼합음성의 평균신호대외곡비는 3.2dB, 표준편차는 3.4dB, 부름말의 평균지속시간은 0.7s이다.

3. 최소분산무외곡응답려파기의 추정

최소분산무외곡응답려파기의 추정에서 부름말이 아닌 배경음성의 공분산행렬을 $\boldsymbol{R}_{nn}$, 부름말의 조향벡토르를 $\boldsymbol{v}$로 표시하면 최소분산무외곡응답려파기 $\boldsymbol{\gamma}$는 다음과 같이 계산된다.

$$\boldsymbol{\gamma}=[\gamma(1),\ \gamma(2),\ \cdots,\ \gamma(K)]^{\mathrm{T}}=\frac{\boldsymbol{R}_{nn}^{-1}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{H}}\boldsymbol{R}_{nn}^{-1}\boldsymbol{v}} \tag{1}$$

여기서 K는 수감부의 개수, 웃첨수기호 T와 H는 각각 행렬의 전위, 공액전위를 표시한다.

$m_{\tau}^{(n)}(k)$를 τ시각에 k번째 통로에서 추정된 비부름말마스크라고 하고 $\overline{m}_{\tau}^{(n)}$을 τ시각에 측정된 비부름말마스크들의 모임

$$\boldsymbol{M}_{\tau}^{(n)}=\left\{m_{\tau}^{(n)}(1),\ m_{\tau}^{(n)}(2),\ \cdots,\ m_{\tau}^{(n)}(c)\right\}$$

의 중위수로 정의하자. 그리고 τ시각에 측정된 다통로진폭스펙트르를

$$\boldsymbol{Y}_{\tau}=[y_{\tau}(1),\ y_{\tau}(2),\ \cdots,\ y_{\tau}(c)]^{\mathrm{T}} \tag{2}$$

로 표기하자.

그러면 배경음성의 공분산행렬 $\boldsymbol{R}_{nn}$은 다음과 같이 추정된다.

$$\boldsymbol{R}_{nn}=\sum_{\tau\in\boldsymbol{T}}\overline{m}_{\tau}^{(n)}\boldsymbol{Y}_{\tau}\left(\overline{m}_{\tau}^{(n)}\boldsymbol{Y}_{\tau}\right)^{\mathrm{H}} \tag{3}$$

여기서 $\boldsymbol{T}$는 부름말구역에 있는 시간프레임첨수들의 모임이다.

부름말의 공분산행렬 $\boldsymbol{R}_{ss}$는 배경음성의 공분산행렬과 류사하게 추정된다.

$$\boldsymbol{R}_{ss}=\sum_{\tau\in\boldsymbol{T}}\overline{m}_{\tau}^{(s)}\boldsymbol{Y}_{\tau}\left(\overline{m}_{\tau}^{(s)}\boldsymbol{Y}_{\tau}\right)^{\mathrm{H}} \tag{4}$$

조향벡토르 $\boldsymbol{v}=[v(1),\ v(2),\ \cdots,\ v(K)]^{\mathrm{T}}$는 공분산행렬 $\boldsymbol{R}_{ss}$를 리용하여 계산한다. 즉 $\boldsymbol{R}_{ss}$의 고유값분해를 진행하고 그것의 최대고유값에 대응하는 고유값벡토르를 $\boldsymbol{v}$의 추정

값으로 결정한다.

추정된 $\boldsymbol{R}_{nn}$과 $\boldsymbol{v}$로부터 식 (1)에 의해 최소분산무외곡응답려파기 $\boldsymbol{\gamma}$를 추정할수 있다.

$\boldsymbol{\gamma}$를 부름말뒤에 오는 지령음성에 해당한 혼합신호 $\boldsymbol{Y}_\tau$에 적용하여 강조된 신호 x_τ를 얻는다.

$$\boldsymbol{x}_\tau = \boldsymbol{\gamma}^h \boldsymbol{Y}_\tau \tag{5}$$

4. 성 능 평 가

혼합신호에서 부름말과 배경음성을 어느 정도로 갈라내는가를 평가하기 위하여 신호대외곡비증가(SDRI)를 리용한다.

신호대외곡비는 목적신호의 진폭스펙트르 $X_{\tau,f}$와 배경신호의 진폭스펙트르 $N_{\tau,f}$, 스펙트로그람 $m_{\tau,f}$에 의하여 다음과 같이 정의된다.

$$SDRI = \frac{1}{\#\boldsymbol{F}} \sum_{f\in\boldsymbol{F}} 10\log_{10}\left(\frac{\sum_{\tau\in\boldsymbol{T}} m_{\tau,f} X_{\tau,f} X^*_{\tau,f}}{\sum_{\tau\in\boldsymbol{T}} m_{\tau,f} N_{\tau,f} N^*_{\tau,f}}\right) - \xi \tag{6}$$

여기서 f는 주파수대역첨수, $\boldsymbol{F}$는 모든 주파수첨수들의 모임, $\#\boldsymbol{F}$는 주파수첨수개수, ξ는 마스크처리를 진행하기 전의 신호대외곡이다.

$$\xi = \frac{1}{\#\boldsymbol{F}} \sum_{f\in\boldsymbol{F}} 10\log_{10}\left(\frac{\sum_{\tau\in\boldsymbol{T}} X_{\tau,f} X^*_{\tau,f}}{\sum_{\tau\in\boldsymbol{T}} N_{\tau,f} N^*_{\tau,f}}\right) \tag{7}$$

우리는 지능고성기의 부름말《무아경》에 대하여 50명의 발성자가 각이한 잡음환경속에서 50번씩 발성한 음성자료를 가지고 심층학습을 진행하여 부름말마스크를 추정하였다. 마찬가지 방법으로 배경음성에 대한 마스크도 추정하였다.

추정의 정확도를 검증하기 위하여 선행한 방법[2]과 제안된 방법의 신호대외곡비증가를 표에 보여주었다.

표. 신호대외곡비

방법	선행방법[2]		제안된 방법	
마스크	부름말	배경음성	부름말	배경음성
신호대외곡비증가	4.9±2.6	3.1±1.6	6.4±1.9	5.8±1.7

표에서는 신호대외곡비증가의 평균과 표준편차를 보여주었다.

표를 통해 알수 있는바와 같이 제안된 방법에서는 신호대외곡비증가가 선행방법보다 보통 1.5dB 높다.

맺 는 말

지능고성기와 같은 음성인식체계들에서 심층학습을 리용한 마스크추정법으로 부름말을 검출한 다음 빔형성려파기를 계산하여 목적발성자의 음성을 강조하는 한가지 방법을

제안하였다. 모의실험을 통하여 제안된 방법이 선행방법보다 배경음성속에서 목적발성자의 음성을 강조하는 성능이 높다는것을 확증하였다.

참 고 문 헌

[1] Hiroshi Sawada et al.; IEEE Trans. Audio, Speech, and Language Processing, 21, 5, 971, 2013.
[2] Marc Delcroix et al.; Proc ICASSP, 5554, 2018.
[3] Takuya Yoshioka et al.; Proc ICASSP, 5739, 2018.

주체108(2019)년 11월 5일 원고접수

Speech Enhancement of Target Speaker Based on Vocative Detection

Ri Ji Un, *Kwak Chong Il*

In speech recognition system such as smart speaker, this method emphasizes the subsequent utterances from target speaker by extracting the vocative with DNN based mask estimator and calculating the beam forming filters.

Keywords: vocative, deep neural network(DNN), blind speech separation(BSS)