한국음향학회지 제25권 제4호 pp. 159-163 (2006)

잡음음성인식을 위한 데이터 기반의 Jacobian 적응방식

A Data-Driven Jacobian Adaptation Method for the Noisy Speech Recognition

정 용 주*
(Chung Young-Joo*)

*계명대학교 전자공학과
(접수일자: 2006년 4월 19일; 채택일자: 2006년 5월 4일)

본 논문에서는 잡음음성인식을 위한 데이터 기반의 향상된 Jacobian 적응 방식을 제안하였다. Jacobian 적응에서 필요로 하는 기준 HMM을 구성하기 위해서 기존에 주로 사용되던 모델결합 방식을 사용하는 대신에 잡음음성을 이용하여 직접 훈련하는 방식을 제안하였다. 이렇게 함으로서 기존의 방법에 비해서 잡음에 의한 음향모델의 변이를 보다 잘 처리할 수 있을 것으로 생각된다. 제안된 방법에서는 Jacobian 행렬의 추정을 위해서 훈련과정에서 Baum-Welch 알고리듬을 사용하였다. 잡음음성에 대한 인식실험을 통해서 제안된 방식이 기존의 Jacobian 적응 방식 뿐 만 아니라 다른 형태의 모델적응 방식들에 비해서도 우수한 성능을 보임을 알 수 있었다.

핵심용어: HMM, 잡음음성인식, Jacobian adaptation

투고분야: 음성처리 분야 (2.5)

In this paper, a data-driven method to improve the performance of the Jacobian adaptation (JA) for the noisy speech recognition is proposed. In stead of constructing the reference HMM by using the model composition method like the parallel model combination (PMC), we propose to train the reference HMM directly with the noisy speech. This was motivated from the idea that the directly trained reference HMM will model the acoustical variations due to the noise better than the composite HMM. For the estimation of the Jacobian matrices, the Baum-Welch algorithm is employed during the training. The recognition experiments have been done to show the improved performance of the proposed method over the Jacobian adaptation as well as other model compensation methods.

Keywords: HMM, Noisy Speech Recognition, Jacobian Adaptation

ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

최근까지 잡음음성인식에서 보다 성능을 높이기 위한 다양한 방법들이 제안되고 개발되었다. 이러한 연구결과는 대체적으로 몇 가지 부류로 구분 지을 수 있는데, 음질향상기법, 잡음에 강인한 특징추출기법 그리고 인식모델 보상방법들로 나누어 질 수 있을 것이다. Hidden Markov model (HMM)에 기반한 인식모델 보상방식에서는 잡음음성으로부터 추출한 잡음의 통계정보를 이용하여 HMM 파라미터 값에 대한 보상이 이루어진다[1-2-3-4]. 특히, JA 방식은 실제 환경과 비슷한 조건에서 HMM을 미리 훈련한 경우에 매우 효과적인 것으로 알려져 있으며 미리 훈련된 HMM (기준 HMM)의 파라미터 값들은 Jacobian 행렬을 이용하여 실제 환경의 잡음음성에 용이하게 적응된다[4]. 기준 HMM의 훈련을 위해서는 일반적으로 parallel model combination (PMC) [1] 이나 NOVO[3] 와 같은 음성과 잡음의 모델결합방식을 주로 이용한다. 이러한 모델 결합방식을 이용함으로서, Jacobian 행렬을 원래의 깨끗한 음성 HMM의 평균벡터값을 이용하여 쉽게 추정할 수 있는 장점이 있다[4]. 하지만, 일반적으로, 모델결합방식을 이용하여 얻어진

책임저자: 정 용 주 (yjjung@kmu.ac.kr)
대구시 달서구 신당동 1000 계명대학교 전자공학부
(전화: 053-580-5925)

HMM은 실제 환경에서 잡음음성을 이용하여 직접 훈련된 HMM에 비해서 그 인식 성능이 다소 떨어진다는 것이 알려져 있다. JA 방식에서는 기준 HMM이 모델결합방식을 통해서 얻어지는데, 이것은 JA 방식이 기존의 PMC 나 NOVO에 비해서 그리 높은 성능을 보이지 못하는 주요 이유가 된다고 생각된다. 본 논문에서는JA 방식의 성능을 향상시키기 위한 방안으로 기준 HMM을 잡음음성을 이용해서 직접 훈련하는 것을 제안하였는데, 이 경우에는 Jacobian 행렬과 깨끗한 음성 HMM 파라미터 간의 관계가 불명료해지므로, Jacobian 행렬을 훈련과정에서 Baum-Welch 방식을 이용하여 추정하였다. 본 논문의 구성은 다음과 같다. 2장에서는 기존의 JA 방식에 대해서 간략히 소개하고 3장에서는 제안된 방식에 대해서 자세히 설명하며 4장에서는 인식실험결과를 소개하고 마지막으로 5장에서 결론을 맺는다.

II. JA 방식의 개요

켑스트럼 (cepstrum) 영역의 잡음음성신호 벡터 $\boldsymbol{y}$는 다음과 같은 비선형식에 의해서 일반적으로 표현될 수 있다[1].

$$\boldsymbol{y} = \boldsymbol{C}\left[\log\{\exp(\boldsymbol{C}^{-1}\boldsymbol{x}) + \exp(\boldsymbol{C}^{-1}\boldsymbol{n})\}\right] \quad (1)$$

여기서 $\boldsymbol{x}$와 $\boldsymbol{n}$ 은 각각 켑스트럼 영역에서의 깨끗한 음성신호와 잡음신호벡터를 나타내며 $\boldsymbol{C}$는 discrete cosine transformation (DCT) 을 나타낸다. 잡음신호 $\boldsymbol{n}$에 대한 잡음음성신호 $\boldsymbol{y}$ 변화율을 나타내는 Jacobian 행렬은 다음과 같이 유도된다.

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{n}} = \boldsymbol{C}\boldsymbol{R}_{\boldsymbol{y}}\boldsymbol{C}^{-1} \quad (2)$$

$$[\boldsymbol{R}_{\boldsymbol{y}}]_{kk} = N_k/(X_k + N_k) \quad (3)$$

$$N_k = (\exp(\boldsymbol{C}^{-1}\boldsymbol{n}))_k \quad (4)$$

$$X_k = (\exp(\boldsymbol{C}^{-1}\boldsymbol{x}))_k \quad (5)$$

식 (3)은 대각 행렬 (diagonal matrix) $\boldsymbol{R}_{\boldsymbol{y}}$의 k 번째의 대각원소 (diagonal element)를 나타낸다. X_k와 N_k는 각각 선형스펙트럼 (linear spectrum) 영역에서의 음성과 잡음벡터의 k번째 원소에 해당한다. Jacobian 행렬의 계산을 위해서는, 먼저 훈련과정에서 미리 가정된 기준잡음신호 (reference noise signal)에 대해서 평균값 $E\{\boldsymbol{n}\}$을 구하고 이를 식(4)의 $\boldsymbol{n}$ 대신에 대입하여 기준잡음신호의 평균에 해당하는 선형스펙트럼 N_k을 구한다. 또한, 연속밀도 HMM의 혼합성분 (mixture component)들에 해당하는 X_k값은 각 혼합성분에 해당하는 깨끗한 음성 HMM의 평균값을 식 (5)의 $\boldsymbol{x}$ 대신에 대입하여 구한다. 이와 같이 얻어진 X_k와 N_k값을 식 (3)에 대입함으로서 기준 HMM의 모든 혼합성분에 대한 Jacobian 행렬을 얻을 수 있게 된다. 얻어진 Jacobian 행렬을 이용하여 아래의 식과 같이 잡음신호가 $\boldsymbol{n}$에서 $\tilde{\boldsymbol{n}}$ 로 변할 경우 잡음음성신호 $\boldsymbol{y}$ 값이 $\tilde{\boldsymbol{y}}$로 변하는 과정을 나타낼 수 있게 된다.

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} + \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{n}}(\boldsymbol{n} - \tilde{\boldsymbol{n}}) \quad (6)$$

잡음음성신호의 평균값을 얻기 위해서는 위식의 양변에 평균자를 적용하여 아래와 같이 구한다.

$$E\{\tilde{\boldsymbol{y}}\} = E\{\boldsymbol{y}\} + \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{n}}(E\{\boldsymbol{n}\} - E\{\tilde{\boldsymbol{n}}\}) \quad (7)$$

III. 데이터기반의 JA 방식

기준 HMM을 모델결합방식을 이용하여 얻는 대신에 본 논문에서는 잡음음성을 이용하여 직접 훈련하는 방안을 제안하였다. 이 경우에는 음성신호와 기준HMM의 혼합성분과의 정렬관계가 불명확해지므로 (모델결합방식에서는 음성신호와 기준 HMM의 혼합성분과의 정렬관계는 기존의 깨끗한 음성 HMM과의 정렬관계를 유지함) 식 (3)~(5)을 이용하여 Jacobian 행렬을 구하기가 어려워진다. 기존의 연구에서[7], 우리는 잡음음성인식에서 HMM 파라미터의 보상을 위하여 해석적인 방법보다는 직접데이터 기반의 추정방식을 사용하는 경우 성능의 향상을 이룰 수 있음을 보였다. 이와 동일한 생각에 기초하여, 본 논문에서는 Jacobian 행렬들을 Baum-Welch 알고리듬[8]에 기반하여 얻고자 한다.

HMM에 기반한 음성인식에서, HMM 파라미터들은 보통 Baum-Welch 알고리듬에 의해 얻어된다. 예를 들면, 연속밀도 HMM의 상태 j 의 혼합성분 k 에 해당하는

평균벡터는 다음과 같은 수식에 의해 추정된다.

$$E(\tilde{y}_t) = \frac{\sum_{t=1}^{T}\gamma_t(j,k)(y_t + \frac{\partial y_t}{\partial n_t}(n_t - \tilde{n}_t))}{\sum_{t=1}^{T}\gamma_t(j,k)} \quad (8)$$

여기서 $\gamma_t(j,k)$는 켑스트럼 특징벡터 x_t가 상태 j의 혼합성분 k에 의해서 발생될 확률을 의미한다.

식 (6)과 (8)을 이용하면, 잡음음성신호 $\tilde{y}_t$ 에 대한 평균벡터는 다음과 같이 추정될 것이다.

$$E(\tilde{y}_t) = \frac{\sum_{t=1}^{T}\gamma_t(j,k)(y_t + \frac{\partial y_t}{\partial n_t}(n_t - \tilde{n}_t))}{\sum_{t=1}^{T}\gamma_t(j,k)} \quad (9)$$

만약, 잡음신호의 차이 $\Delta n(=n_t - \tilde{n}_t)$ 값이 평균치 (즉, 시간 t 에 독립적)로서 대체가 가능하다면, 위의 식 (9)는 다음과 같이 전개될 수 있다.

$$E(\tilde{y}_t) = \frac{\sum_{t=1}^{T}\gamma_t(j,k)y_t}{\sum_{t=1}^{T}\gamma_t(j,k)} + \frac{\sum_{t=1}^{T}\gamma_t(j,k)\frac{\partial y_t}{\partial n_t}}{\sum_{t=1}^{T}\gamma_t(j,k)}\Delta n$$
$$= E(y_t) + \frac{\sum_{t=1}^{T}\gamma_t(j,k)(CR_{y_t}C^{-1})}{\sum_{t=1}^{T}\gamma_t(j,k)}\Delta n \quad (10)$$

$$\mu_{\tilde{y}} = E(y_t) + E(CR_{y_t}C^{-1})\Delta n = \mu_y + \mu_J \Delta n \quad (11)$$

여기서 μ_y는 기준 HMM의 평균벡터를 의미하고 μ_J는 추정된 Jacobian 행렬이다. Δn 은 훈련시의 기준잡음신호의 평균값과 인식시의 관측 잡음신호의 평균값의 차이를 구하여 얻어진다. 훈련과정을 통해서 μ_y 와 μ_J 를 먼저 구한 후에 인식과정에서 Δn 을 이용하여 최종적인 잡음음성에 대한 평균값 $\mu_{\tilde{y}}$을 식 (11)에 의해 얻게 된다. 제안된 방식의 차별점은 Jacobian 행렬 μ_J 가 기준HMM의 다른 파라미터들과 마찬가지로 잡음음성을 이용한 EM 과정을 통해서 추정된다는 점이다.

IV. 인식 실험 결과

잡음환경에서 화자독립 단어 인식실험을 통해서 제안된 방식의 성능을 평가하였다. 인식대상 어휘는 음소분포가 비교적 고르게 되어 있는 한국어 75 단어이며 음향모델을 위한 기본단위는 32개의 유사음소를 사용하였다. 각각의 유사음소단위는 연속밀도 HMM에 의해서 모델링된다. 화자의 수는 80명이며 이들은 각각 75단어를 한번씩 발성하였다. 인식실험을 위해서 잭-나이프 (Jack-knife) 방식을 이용하였다. 전체 화자를 20명씩 4개의 그룹으로 나눈 후, 그 중 하나의 그룹은 인식용으로 나머지 3그룹은 훈련용으로 활용하였다. 이와 같은 과정을 4회 반복하여 인식실험을 수행하여 인식화자의 수를 4배로 증가시키는 효과를 거두도록 하였다. 잡음음성을 얻기 위해서는 원래의 깨끗한 음성에 차량잡음과 배블(babble)잡음을 다양한 신호대잡음비에 맞추어 더해 주었다. 잡음신호는 AURORA 2 데이터에 있는 잡음파일로부터 얻었다. 인식특징벡터로는 13차의 멜주파 (mel-frequency) 켑스트럼 계수 (MFCC)와 그의 차분계수 (delta-MFCC)를 사용하였다. 표1과 2에는 제안된 방식, 즉 데이터기반의 JA (D-JA)에 대한 인식결과가 나타나 있다. 표에 나타난 인식율은 MFCC의 정적 평균벡터만을 보상한 경우에 해당한다. 표의 결과에서 보면 PMC와 JA 방식은 큰 성능의 차이를 보이지 않음을 볼 수 있다. PMC 과정에서 잡음통계정보를 추정하기 위한 묵음구간을 음성의 앞부분의 20프레임 (0.2초)으로 잡았는데, 잡음의 특성을 파악하기에 비교적 충분한 통계치를 얻을

표 1. 자동차 잡음환경에서 제안된 방법(D-JA)과 기존 방식간의 단어인식율의 비교

Table 1. Comparison in word recognition rates (%) of the proposed method (D-JA) with previous methods in the noisy speech recognition (Car noise).

	0 dB	10 dB	20 dB	Clean
Baseline	12.6	60.7	92.5	98.6
Re-training	82.1	95.0	97.5	98.6
PMC	59.8	87.8	95.3	98.6
JA	59.9	87.8	95.4	98.6
D-JA	82.4	95.0	97.4	98.6

표 2. 배블 잡음환경에서 제안된 방법(D-JA)과 기존 방식간의 단어인식율의 비교

Table 2. Comparison in word recognition rates (%) of the proposed method (D-JA) with previous methods in the noisy speech recognition (Babble noise).

	0 dB	10 dB	20 dB	Clean
Baseline	13.6	61.7	92.5	98.6
Re-training	79.7	93.9	96.8	98.6
PMC	54.9	86.6	95.1	98.6
JA	54.2	86.7	95.3	98.6
D-JA	79.7	94.0	96.9	98.6

표 3. 배블 잡음환경에서 인식시의 SNR값이 변하는 경우의 제안된 D-JA와 다른 모델 변환 방식 및 재훈련 방법간의 인식율의 상호비교

Table 3. Performance comparison in word recognition rates (%) of the proposed method (D-JA) with other model compensation methods and the re-training method when the SNR of the testing speech changes. (Babble noise).

	훈련 \ 인식	0dB	5dB	15dB	25dB
D-JA	10 dB	79.7	90.1	95.2	93.4
	20 dB	64.5	85.2	95.3	97.7
JA	10 dB	56.1	74.8	92.0	95.8
	20 dB	58.0	77.1	92.4	96.8
Re-training	10 dB	71.3	89.4	94.8	92.7
	20 dB	31.6	69.7	95.3	97.4
PMC		54.9	74.2	92.1	96.9

수 있었다고 생각된다. 하지만 실제에 있어서 묵음구간의 길이가 충분치 못한 경우에는 PMC 성능이 다소 떨어질 것으로 예상된다.

제안된 D-JA 방법은 PMC 나 JA 방식에 비해서 모든 신호대잡음비 (SNR: Signal to noise ratio)에서 향상된 인식성능을 보임을 알 수 있었다. 이는 기준 HMM을 잡음음성을 이용하여 직접적으로 훈련한데 그 주요 원인이 있다고 생각된다. 제안된 D-JA 방식은 PMC 나 JA 방식에 비해서 인식에러의 상대적인 감소율이 잡음음성인식에서 약 40~50(%) 에 이르는 우수한 성능을 보였다.

표1과 2에서의 우수한 성능을 얻기 위해서는 각각의 SNR에 대해서 독자적인 기준 HMM을 가지고 있어야 한다. 하지만 많은 종류의 기준 HMM을 인식시에 가지고 있어야 하는 것이 실제적으로 어렵기 때문에 우리는 제안된 방식의 강인성을 검토비교 하였다. 이를 위해서는 특정한 SNR 조건하에서 훈련된 기준 HMM을 이용하여 다양한 SNR을 가진 인식환경에서 실험하였다. 표3에는 기준HMM이 각각 10 dB과 20 dB의 배블 잡음 환경에서 훈련된 경우에 인식성능을 보여준다. 기대한 데로 D-JA 방식은 기존의 방식에 비해서 우수한 강인성을 보여 주었다. 예를 들어 기준HMM이 10 dB에서 훈련된 경우에 D-JA 방식은 0 dB에서 79.7(%)의 인식율을 나타낸 반면 기존의 JA 방식은 동일한 조건하에서 56.1(%)의 인식성능을 보임을 알 수 있다. 또한 20 dB에서 기준 HMM을 훈련한 경우 D-JA 방식은 PMC 나 재훈련의 경우에 비해서도 향상된 성능을 보임을 알 수 있었다. 기존의 JA 방식은 인식환경이 높은 SNR 인 경우에는 다소 좋은 성능을 보이기도 하지만 전반적으로 제안된 방식에 비해서는 다양한 SNR 에 대한 강인성이 저조 한 것으로 보여 진다. 이와 같은 D-JA 방식의 다양한 SNR에 대한 강인성은 소수의 기준 HMM만을 제공할 수 있는 실제 환경에서 매우 효과적일 것이라고 생각된다.

V. 결 론

본 논문에서는 보다 향상된 JA을 위해서 기준 HMM을 잡음음성으로 직접 훈련하는 방안을 제안하였다. 모델결합방식에 비해서 제안된 방식은 보다 강인한 기준HMM을 구성할 수 있었으며, 기존의 JA 방식뿐만 아니라 기타의 모델 보상방식에 비해서도 우수한 인식성능을 보임을 알 수 있었다. 제안된 방식의 강인성은 특히 소수의 기준HMM 만이 사용가능한 실제 인식 환경에서 매우 유용할 것으로 생각된다.

감사의 글

본 연구는 산업자원부 지역연구개발클러스터구축사업 중 경북대학교 첨단 진단/예측 의료기술 클러스터 사업단의 연구비지원을 받아 수행되었음.

참 고 문 헌

1. Gales, M.J.F., *Model based techniques for robust-speech recognition*, (Ph. D. Dissertation, University of Cambridge, 1995)
2. Moreno, P.J., *Speech recognition in noisy environments*, (Ph. D., Dissertation, Carnegie Mellon University 1996)
3. Martin, F., Shikano, K. and Minami, Y., "Recognition of noisy speech by composition of hidden Markov models" In proc. Eurospeech 93, 031-1034, 1993.
4. Sagayama, S., Yamaguchi, Y. and Takahashi, S., "Jacobian adaptation of noisy speech models", IEEE Workshop on *Automatic Speech Recognition and Understanding*, Dec. 1997, 396-403.
5. Hung, J.W., Shen, J.L. and Lee, L.S., "New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination (PMC) techniques" IEEE Trans. Speech and Audio Processing, 2001, 9(8), 842-855.
6. Moreno, P.J., Raj, B. and Stern, R.M., "Multivariate Gaussian-Based Cepstral normalization", In Proc. ICASSP 95, 1995.
7. Chung, Y.J., "A data-driven approach for the model parameter compensation in noisy speech recognition," In

proc. Interspeech 2005, Lisboa, 2005, 961-964.
8. Baum, L.E., Petrie, G.S.T. and Weiss, N. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math.Statist., **41** 164-171, 1970.

저자 약력

• **정 용 주 (Chung Young-Joo)**

1988년 서울대학교 전자공학과 (학사)
1990년 한국과학기술원 전기전자공학과 (석사)
1995년 한국과학기술원 전기전자공학과 (박사)
1999년~현재: 계명대학교 전자공학과 교수
※주관심분야: 음성인식, 패턴인식, 통계신호처리