은닉 마르코프 모델의 최대 마진 훈련을 이용한 음성 감정 인식

A Max-Margin Learning of Hidden Markov Model for Speech Emotion Recognition

저자 (Authors) 윤성락, 이동훈, 백승렬, 박상혁, 장달원, 유창동
Sungrack Yun, Donghoon Lee, Seungryul Baek, Sanghyuk Park, Dalwon Jang, Chag D. Yoo

출처 (Source) 대한전자공학회 학술대회 , 2010.6, 4-7 (4 pages)

발행처 (Publisher) 대한전자공학회
THE INSTITUTE OF ELECTRONICS ENGINEERS OF KOREA

URL http://www.dbpia.co.kr/Article/NODE02347124

APA Style 윤성락, 이동훈, 백승렬, 박상혁, 장달원, 유창동 (2010). 은닉 마르코프 모델의 최대 마진 훈련을 이용한 음성 감정 인식. 대한전자공학회 학술대회, 4-7.

은닉 마르코프 모델의 최대 마진 훈련을 이용한 음성 감정 인식

*윤성락, 이동훈, 백승렬, 박상혁, 장달원, 유창동
한국과학기술원 전기 및 전자공학과
e-mail : *yunsungrack@kaist.ac.kr, dongfu@kaist.ac.kr, korealovewe@naver.com, mudge@kaist.ac.kr, dal1@kaist.ac.kr, cdyoo@ee.kaist.ac.kr*

A Max-Margin Learning of Hidden Markov Model for Speech Emotion Recognition

*Sungrack Yun, Donghoon Lee, Sanghyuk Park, Seungryul Baek, Dalwon Jang, Chag D. Yoo
Department of Electrical Engineering
Korea Advanced Instituted of Science and Technology

Abstract

In this paper, we propose a max-margin learning algorithm of hidden Markov model for speech emotion recognition. A max-margin learning leads to a good generalization ability on testing data even with small number of training data which may lead to an over-fitting. In the experiment, we observed that the proposed learning algorithm outperforms the learning criteria such as the maximum likelihood and maximum mutual information.

I. 서론

인간과 로봇의 상호작용 기술[1]에 대한 연구가 최근 다양한 면에서 많이 진행되어 오면서 인공지능 기술이 크게 발전하였다. 그 중 음성/화자 인식 기술은 오래된 연구 분야 중의 하나로서, 최근 대부분의 인공지능 로봇이나 시스템에는 음성/화자 인식기가 들어있다. 따라서, '무슨 말'을 '누가' 했는지 시스템이 알 수 있으며, 이에 따라 적절한 반응을 할 수 있게 된다. 더 나아가, 만약 로봇이 감정을 인식할 수 있다면, '무슨 말'을 '누가'했는지 뿐만 아니라, '어떻게' 말을 하고 있는지 알 수 있게 된다. 이러한 감정 인식은 인간과 로봇의 상호작용에서 로봇이 좀 더 인간에 가까운 적절한 반응을 할 수 있도록 큰 역할을 한다.

인간의 감정은 음성, 표정, 행동 등 여러 가지 형태로 다양하게 나타나는데, 이 논문에서는 음성을 이용한 감정 인식에 초점을 둔다. 인간의 감정 종류에는 화남, 행복, 기쁨, 슬픔, 우울, 실망, 두려움, 놀람, 의심, 중립 등 여러 가지가 있는데, 여기서는 크게 다섯 가지의 감정, 즉, 화남, 행복, 슬픔, 중립, 놀람을 구별하는 것을 목표로 한다.

음성 감정 특징점으로는 음조(pitch), 음조의 범위(pitch range), 음성의 발화 속도(speech rate), 로그 에너지, 포먼트(formant), 멜-밴드(mel-band)의 에너지, MFCC(mel frequency cepstral coefficients) 등을 많이 사용한다. 인식 성능을 높이고, 계산량을 줄이기 위해서 특징점 선택(feature-selection)]을 할 수도 있는데 [2], 이 논문에서는 MFCC만을 감정 특징으로 이용하고, 감정 모델의 학습방법의 개선을 통해 감정 인식률을 향상시키고자 한다.

가장 많이 사용되는 감정 모델로는 은닉 마르코프 모델 (Hidden Markov Model, HMM)이 있다. HMM의

파라미터 추정 방법으로서 최대 우도 (maximum likelihood, ML) 기법 [3] 이 가장 많이 사용된다. 최대 우도 기법은 훈련 데이터 숫자가 무한대에 가까울 정도로 많고, 묘사하려는 특징점이 HMM을 잘 따른다면, 우수한 성능을 낸다는 것이 [4]에 증명되어 있다. 하지만, 실제로 데이터의 개수는 한정적이며, 또한 특징점들이 정확히 HMM을 따르는 것도 아니다. 이러한 경우에는 변별력 훈련 (discriminative training) 기법이 더 우수한 성능을 나타낸다.

가장 많이 쓰이는 변별력 훈련 기법으로는 최대 상호 정보 기법 (maximum mutual information, MMI)이 있다 [5]. 하지만, MMI는 훈련데이터의 인식오류를 직접 줄이는 방법으로써, 테스트 데이터의 인식오류를 직접 줄이지는 못한다. 테스트 데이터의 인식오류를 직접 줄이기 위해서는 최대 마진 기법을 사용하여야 한다 [6]. 이 논문에서는 HMM의 훈련을 최대 마진 기법을 사용하여 함으로써, 테스트 감정 데이터의 인식 오류를 줄이고자 한다.

실험은 Danish Emotional Speech (DES)를 사용하여 감정 인식을 수행하였으며, 실험결과에서 제안한 훈련 방법이 더 우수함을 확인하였다.

II. 배경

2.1 HMM을 이용한 감정 인식

이 논문에서는, 음성이 하나의 감정으로 발화되었다고 가정하고, 따라서 하나의 상태를 갖는 HMM으로 감정을 모델링하였다. 음성 특징 벡터 집합 $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$는 음성의 첫 번째 프레임에서 T 번째 프레임까지 추출된 특징 벡터의 모임이며, 이를 이용하여 HMM의 파라미터를 훈련한다. 감정 인식은 다음의 기준에 의하여 이루어진다.

$$\mathbf{y}^* = \arg\max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{X}, \mathbf{y}; \theta)$$

즉, 감정 인식은 총 M개의 감정 레이블 $\mathbf{y}$에 대하여 discriminant function $F(\mathbf{X}, \mathbf{y}; \theta)$ 를 최대화 하는 레이블을 찾는 것이다. Discriminant function의 파라미터, 즉 HMM의 파라미터를 θ로 가정하였다. HMM에서는 discriminant function 으로 다음의 확률을 사용한다.

$$F(\mathbf{X}, \mathbf{y}; \theta) = \log p_\theta(\mathbf{X}|\mathbf{y})p(\mathbf{y})$$

여기서, $p(\mathbf{y})$는 사전 확률로써 모든 감정 레이블에 대해 동일하다고 가정하였다. 그리고 HMM의 상태 아웃풋 확률은 K 개의 가우시안 혼합 확률로 모델링하여 아래와 같이 쓸 수 있다.

$$\begin{aligned} F(\mathbf{X}, \mathbf{y}; \theta) &= \log p_\theta(\mathbf{X}|\mathbf{y})p(\mathbf{y}) \\ &= \log\left[\frac{1}{M}\prod_{t=1}^{T}\sum_{k=1}^{K} w_k N(\mathbf{x}_t; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)p(\mathbf{y})\right] \end{aligned}$$

2.3 최대 마진 훈련

최대 마진 기법 [6, 7]을 이용한 HMM의 최대 마진 훈련 기준은 다음과 같이 쓸 수 있다.

$$\begin{aligned} \min_{\rho,\boldsymbol{\xi},\theta:||\theta||=\gamma} \quad & -\rho + \frac{C}{N}\sum_{n=1}^{N}\xi_n \\ \text{subject to} \quad & d(\mathbf{X}_n, \mathbf{y}; \theta) \geq \rho - \xi_n, \ \forall n \\ & \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}_n, \ \rho \geq 0, \ \xi_n \geq 0, \ \forall n \end{aligned}$$

여기서, $d(\mathbf{X}_n, \mathbf{y}; \theta) = F(\mathbf{X}_n, \mathbf{y}_n; \theta) - F(\mathbf{X}_n, \mathbf{y}; \theta)$ 이다. 위 기준은, 정답 감정 레이블이 주어졌을 때의 discriminant function 과 그렇지 않은 레이블이 주어졌을 때의 discriminant function의 차이를 최대화 할 수 있도록 하는 훈련이다. 이 때, ρ는 두 function의 차이 즉 마진이라고 한다. ξ_n은 정답 감정 레이블이 주어졌을 때의 discriminant function 과 그렇지 않은 레이블이 주어졌을 때의 discriminant function의 차이를 항상 양수 값으로 할 수 없을 때, 즉, 훈련데이터의 인식 오류를 허용하기 위하여 필요한 슬랙 변수이다. 인식오류의 최소화와 마진의 최대화는 파라미터 C에 의해 조절된다. 위를 최적화하는 파라미터 θ를 찾는 것이 제안된 훈련 방법이다.

감정 인식 데이터의 경우에는 사람에 따라 똑같은 감정을 발화하는 특징이 많이 다르고, 또한 정확히 발화한 데이터의 수집에 어려움이 있어, 감정 모델이 훈련 데이터에 오버 피팅 되는 경우가 많이 발생한다. 이러한 경우에는 ML훈련보다 위에서 제안된 최대 마진 훈련 기법이 좋은 성능을 낸다.

특히, 감정 인식의 경우에 최대 마진 기법을 적용하기 수월하다. 음성 인식과 같은 경우에는 레이블이 연속된 단어일 수 있는데, 그러한 경우에는 정답이 아닌 레이블들이 아주 많이 발생하기 때문에, margin constraint의 수가 매우 많아져 계산량이 많아지는 문제점이 있다. 하지만 감정 인식의 경우, 구별해야 할 감정의 개수가 적고, 한 음성은 하나의 감정으로 발화하기 때문에, 정답이 아닌 레이블의 수가 그렇게 많지

않다. 따라서 마진 constraint를 줄여야 할 번거로움이 없다.

III. 최적화

이 논문에서 제시한 최대 마진 훈련 기준은 semi-definite programming을 사용하여 최적화를 수행하였다. Semi-definite programming을 이용하여 슬랙 변수가 없는 경우에 HMM을 훈련하는 방법이 [8]에서 제안되었다. 이 논문에서는 슬랙 변수를 사용한 경우로 확장하여 semi-definite programming을 하여 제안한 기준을 최적화 하였으며, semi-definite programming solver로는 DSDP[9]를 사용하였다.

IV. 실험

감정 인식 데이터베이스로 DES를 사용하였다. 두 명의 남자와 여자 배우가 5가지의 감정(화남, 행복, 중립, 슬픔, 놀람)을 발화하여 수집된 데이터베이스이다. 각 배우는 13개의 문장을 각 감정에 따라 발화하였다. 그리고 추가로 81개의 중립 감정의 음성이 수집되어 총 341개의 데이터가 수집되었다. 여성으로부터 총 175개, 남성으로부터 총 166개가 수집되었다.

데이터베이스는 총 네 개의 그룹으로 나뉘어져, 두 그룹의 데이터는 훈련 데이터, 나머지 한 그룹 데이터는 테스트 데이터, 그리고 나머지 한 그룹의 데이터는 개발 데이터로 사용하였다. 개발 데이터는 파라미터 C를 찾기 위하여 사용된다.

음성 특징은 12차원의 MFCC와 로그 에너지 그리고 이들의 delta, acceleration 값을 추가하여 총 39차원의 특징 벡터를 사용하였다. 25ms의 길이의 음성 프레임을 사용하였고, 음성 프레임은 10ms의 길이로 이동하면서 특징 벡터를 추출하였다.

가우시안 개수	ML(%)	MMI(%)	최대 마진 훈련(%)
1	26.40	29.55	35.82
2	43.57	48.80	57.35
4	50.38	56.05	65.34

표 1 가우시안 개수를 1에서 4까지 변화시켜 가며 얻은 감정 인식률

서로 다른 개수의 가우시안을 사용하였을 때, 그 인식결과를 표1에 나타내었다. 우선 ML모델을 얻고, 이 ML모델을 초기값으로 하여 MMI 훈련과 최대 마진 훈련을 수행하였다. MMI 훈련은 ML보다 조금 우수한 성능을 나타내었다.

최대 마진 훈련은 MMI 보다 훨씬 더 우수한 성능을 나타내었다. DES의 적은 데이터 수에 의해 모델 파라미터가 오버 피팅되었기 때문에, 최대 마진 훈련이 ML MMI보다 우수한 성능을 나타내게 된다. 최대 인식률을 가우시안을 네 개를 사용하였을 때, 65.34%였다.

표 2에는 최대 마진 훈련에서, 각 감정에 따른 인식률을 표시하였다. 이 표를 보면 중립이라는 감정을 인식하는 데에 오류가 가장 적음을 알 수 있고, 놀람이라는 감정을 인식하는 데에 오류가 가장 큼을 알 수 있다.

	화남	슬픔	중립	기쁨	놀람
화남	57.69	0	19.23	9.62	13.46
슬픔	0	51.92	40.39	1.92	5.77
중립	0	10.67	88.77	0	0.56
기쁨	15.38	11.54	17.31	44.23	11.54
놀람	17.31	13.46	28.85	11.54	28.85

표 2 최대 마진 훈련에서 각 감정에 따른 인식률(%)

V. 결론 및 향후 연구 방향

본 논문에서는 음성 감정 인식을 위한 HMM의 최대 마진 훈련 기법에 대해 제안하였다. 최대 마진 기법은 테스트 데이터의 오류를 직접 줄이는 방법으로서, ML 훈련으로 얻은 HMM보다 그 성능이 우수하다. 특히, 훈련데이터가 적고, 모델 파라미터가 훈련 데이터에 오버 피팅 된 경우, 우수한 성능을 나타낸다. 실험에서는 DES 데이터 베이스를 사용하여 ML 훈련, MMI 훈련, 최대 마진 훈련의 성능을 비교하여 보았다. 실험 결과로부터, 최대 마진 훈련이 나머지 두 훈련 방법보다 우수함을 알 수 있었다.

참고문헌

[1] R. W. Picard, Affective Computing, MIT Press, 1997.

[2] D. Ververidis and C. Kotropoulos, “Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition,” Signal Processing,

vol. 88, no. 12, pp. 2869 - 3014, 2008.

[3] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in Proceedings of the IEEE, vol. 77, no. 2, 1989, pp. 257 - 286.

[4] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional vesus conditional maximum likelihood," IEEE Trans. Acoust., Speech, Signal Processing, vol. 31, no. 4, pp. 814 - 817, 1983.

[5] A. B. Yishai and D. Burshtein, "A discriminative training algorithm for hidden Markov models," IEEE Trans. Speech Audio Processing, vol. 12, no. 3, pp. 204 - 217, 2004.

[6] I. Tsochantaridis, T. Joachims, and T. Hofmann, "Large margin methods for structured and interdependent output variables," Journal of Machine Learning Research, vol. 6, pp. 1453 - 1484, 2005.

[7] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," Advances in Neural Information Processing Systems, vol. 16, 2004.

[8] Y.Yin and H. Jiang, "A compact semidefinite programming (SDP) formulation for large margin estimation of HMMS in speech recognition," in Proc. ASRU, 2007, pp. 312-317.

[9] S. J. Benson and Y. Ye, "Algorithm 875: DSDP5—software for semidefinite programming," A Research Journal of the Association for Computing Machinery on Mathematical Software, vol. 34, no. 3, pp. 16:1-16-20, 2008.