GMM supervector linear kernel SVM을 이용한 화자식별에서 학습 데이터 길이와 개수간의 상관관계 연구

소 병 민*, 김 경 화**, 김 민 석***, 양 일 호*, 김 명 재*, 유 하 진*
*서울시립대학교 컴퓨터과학부
**대검찰청 음성분석실
***LG전자 전자기술원

A Research on the Correlation between the Lengths of Training Utterances and the Number of Training Data in Using GMM supervector linear kernel SVM for Speaker Recognition

Byung-Min So*, Kyung Wha Kim**, Min-Seok Kim***, IL-Ho Yang*,
Myung-Jae Kim*, Ha-Jin Yu*
*School of Computer Science, University of Seoul
**Supreme Prosecutor's Office
***Advanced Research Institute, LG Electronics
sbm1210@naver.com, savoix@spo.go.kr, ms@uos.ac.kr, heisco@hanmail.net,
arthmody@naver.com, hjyu@uos.ac.kr

Abstract

We have been building a speaker identification system using GMM (Gaussian mixture model) supervector linear kernel SVM (support vector machine), which is one of the state-of-the-art methods studied so far. We have shown that by dividing a training utterance into a small number of tokens we could improve the performance. In this research, we investigated the relation between the utterance length and the number of segments to be used for the SVM supervectors.

1. 서론

GMM(Gaussian mixture model)[1] supervector를 특징으로 사용한 SVM(support vector machine)[2]은 하나의 발성으로부터 하나의 특징 벡터를 추출한다. 그러므로 여러 개의 발성을 사용해야 많은 수의 특징 벡터를 추출할 수 있다. 많은 수의 특징 벡터는 SVM 학습에 사용되는 서포트 벡터(support vector)의 선택에 다양성을 주고 결과적으로 SVM의 분류 성능에 영향을 미친다. 그리고 소수의 학습데이터가 주어졌을 때 기존데이터를 사용한 것보다 더 많은 특징 벡터를 추출하기 위하여 학습데이터를 분할하는 것으로 SVM의 분류 성능을 향상시킬 수 있다[3]. 하지만 학습데이터를 너무 짧은 길이로 분할하는 경우 인식률이 하락하는 경향을 보였다. 따라서 본 논문에서는 학습데이터의 길이와 분할 개수간의 상관관계에 대해서 연구하였다.

2. DB

본 연구에서는 2008년도에 대검찰청에서 수집된 쌍둥이 음성 DB를 사용하였다. 쌍둥이 음성 DB는 일란성 남자 쌍둥이 16쌍, 여자 쌍둥이 17쌍, 이란성 남자 쌍둥이 3쌍, 여자 쌍둥이 5쌍(총 일란성 33쌍, 이란성 8쌍의 쌍둥이)으로 구성되어 있다. 발성한 내용은 두 종류의 문단을 각각 3회 반복 발성한 낭독체 음성과 자유발화 형식의 인터뷰이다.

3. 특징 추출

본 연구에서는 UBM(universal background model)[4] 생성을 위해 15차 MFCC(Mel-frequency cepstral coeffients)에 로그 에너지 더한 16차 특징을 사용하였다. 그리고 1024개의 mixture를 갖는 UBM에 학습데이터를 MAP(maximum a posteriori)[5] 적응하여 GMM 평균들의 집합인 GMM supervector를 추출하였다.

4. 화자식별 실험 및 결과

본 연구에서는 화자식별 실험에 쌍둥이 음성 DB의 전체 인원 중 남녀 일란성 쌍둥이 30쌍(남자 15쌍, 여자 15쌍) 음성 데이터를 사용하였다. 그리고 ETRI에서 수집한 한국어 중가마이크 DB의 주차 데이터 중 남녀 100명(남자 50명, 여자 50명)을 사용하여 UBM을 생성하였다. 각각의 DB는 8kHz로 샘플링되었다. 표1에 보이는 것과 같이 학습데이터의 길이와 분할 개수에 따른 경향성을 파악하기 위하여 하나의 모델에 사용되는 학습데이터의 길이(가로축)를 10초부터 70초까지 10초씩 늘려가며 실험을 진행하였다. 그리고 하나의 모델당 분할되는 학습데이터의 수(세로축)는 1부터 30까지 1씩 증가시켜 실험을 진행하였다.

표 1. 쌍둥이 DB 학습데이터 분할에 따른 인식률

	10초	20초	30초	40초	50초	60초	70초
1	59.3	63.4	62.2	64.1	62.9	66.4	64.1
2	78.9	83.8	80.1	80.3	80.8	80.5	78.0
3	86.3	90.8	90.8	88.8	86.5	86.5	84.4
4	89.0	92.4	93.1	92.0	90.8	89.7	88.1
5	89.0	94.1	94.3	93.8	93.1	92.9	89.9
6	**89.5**	94.3	95.4	95.4	94.7	94.1	93.4
7	**89.5**	94.7	96.1	95.7	94.7	95.2	94.7
8	89.0	94.3	95.9	95.4	95.2	96.1	95.9
9	87.2	95.0	96.6	96.1	95.7	96.6	95.7
10	87.9	95.0	96.6	95.9	95.4	96.1	96.1
11	87.9	**95.4**	96.1	96.1	96.3	97.3	96.3
12	88.1	93.8	96.1	**96.6**	**97.0**	97.0	97.0
13	88.6	94.5	96.8	96.1	96.6	**97.5**	96.8
14	86.5	94.7	**97.0**	95.9	96.3	97.3	**97.5**
15	86.0	94.1	95.9	96.1	96.6	97.0	96.8
16	86.5	93.4	96.1	96.1	96.3	**97.5**	97.0
17	87.0	95.0	96.1	96.3	96.8	97.0	96.6
18	85.4	92.9	95.9	96.1	96.8	**97.5**	96.6
19	85.6	93.1	96.1	96.1	96.8	96.8	**97.5**
20	84.9	91.5	95.9	95.7	96.8	96.8	97.3
21	87.2	92.4	95.9	95.4	96.8	96.8	97.3
22	85.6	92.7	95.9	96.3	96.6	97.3	97.0
23	86.3	91.8	96.3	95.9	96.6	**97.5**	96.6
24	86.0	90.8	95.2	95.7	**97.0**	96.8	97.0
25	84.2	91.5	95.2	96.1	96.3	96.3	96.8
26	85.4	91.1	96.6	95.7	96.6	96.8	96.6
27	82.8	92.0	94.7	95.4	96.6	96.6	97.3
28	85.1	91.5	95.9	95.4	96.8	96.6	96.8
29	81.5	89.7	95.4	96.1	96.8	96.8	96.3
30	84.4	89.5	95.2	95.2	96.3	96.3	96.6

실험결과, 하나의 모델에 사용되는 학습데이터의 길이가 늘어날수록 높은 인식률을 보이는 분할 개수도 커진다는 것을 알 수 있었다.

5. 결론

본 연구에서는 SVM을 사용한 화자인식에서 학습데이터의 길이와 분할 개수간의 상관관계에 대하여 연구하였다. 쌍둥이 음성 DB를 사용한 실험을 통하여 학습데이터의 길이 L에 대한 높은 인식률을 보이는 분할 개수 K가 다음 식과 같은 경향성을 갖는 것을 알 수 있었다.

$$K = a + bL \qquad (1)$$

위의 식에서 a와 b는 상수를 나타낸다. 향후 다양한 데이터베이스를 사용한 추가 실험으로 상수 a와 b의 평균적인 값을 추정하는 연구를 하고자 한다.

Acknowledgement

본 논문은 2011년도 대검찰청 연구용역의 지원으로 수행된 연구입니다. (과제명: 녹음채널별 화자 자동식별시스템 개발)

참고문헌

[1] Douglas A. Reynolds and Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. Speech Audio Processing*, Vol.3, No.1, pp.72-83, 1995.

[2] W.M. Campbell, D.E. Sturim, D.A. Reynolds "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, Vol.13, pp.308-311, 2006.

[3] 소병민, 김경화, 김민석, 양일호, 김명재, 유하진, “특징 강화 기법과 학습 데이터 길이 조절에 의한 Supervector Linear Kernel SVM 화자식별 개선,” 한국음향학회지, Vol.30, No.6, pp. 330-336 , 2011.

[4] Douglas A. Reynolds, Thomas F. Quatieri and Robert B. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing., Vol.10, pp.19-41, Jan. 2000.

[5] J.-L. Gauvain and C.-H. Lee. "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech Audio Proc.*, Vol.2, pp.291-298, Apr. 1994.