논문 2011-48SP-2-17

SVM의 미세조정을 통한 음성/음악 분류 성능향상

(Fine-tuning SVM for Enhancing Speech/Music Classification)

임 정 수*, 송 지 현**, 장 준 혁***

(Chungsoo Lim, Ji-Hyun Song, and Joon-Hyuk Chang)

요 약

Support vector machine (SVM)은 패턴인식 분야에 많이 사용되어지고 있다. 한 예로서 3GPP2 selectable mode vocoder (SMV)와 같은 규격화된 코덱에 쓰여 코덱의 음성/음악 분류 성능을 향상시킬 수 있다. 본 논문에서는 SVM을 개선시켜 음성/음악의 분류성능을 향상시키는 새로운 방법을 제안한다. SVM을 학습시킬 때 적용되는 기존의 기법들과는 달리 제안되는 기법은 SVM이 패턴분류를 행할 때 사용된다. 그렇기 때문에 기존의 기법들과 독립적으로 개발되고 사용될 수 있고, 따라서 패턴분류의 성능을 한층 더 향상시킬 수 있다. 이를 위해 먼저 radial basis function의 커널 width 파라미터가 SVM의 패턴분류에 미치는 영향을 분석해 보았다. 분석한 결과, 커널 width 파라미터를 가지고 SVM의 패턴분류 성향을 미세 조정할 수 있다는 것을 알았다. 또한 음성신호의 각 프레임 간의 상관관계 (correlation)을 확인하고 이를 커널 width 파라미터조절의 길잡이로 삼았다. 실험을 통해, 제안된 기법이 SVM의 성능을 향상시킬 수 있음을 증명하였다.

Abstract

Support vector machines have been extensively studied and utilized in pattern recognition area for years. One of interesting applications of this technique is music/speech classification for a standardized codec such as 3GPP2 selectable mode vocoder. In this paper, we propose a novel approach that improves the speech/music classification of support vector machines. While conventional support vector machine optimization techniques apply during training phase, the proposed technique can be adopted in classification phase. In this regard, the proposed approach can be developed and employed in parallel with conventional optimizations, resulting in synergistic boost in classification performance. We first analyze the impact of kernel width parameter on the classifications made by support vector machines. From this analysis, we observe that we can fine-tune outputs of support vector machines with the kernel width parameter. To make the most of this capability, we identify strong correlation among neighboring input frames, and use this correlation information as a guide to adjusting kernel width parameter. According to the experimental results, the proposed algorithm is found to have potential for improving the performance of support vector machines.

Keywords: Support Vector Machine (SVM), Selectable Mode Vocoder (SMV), Kernel, Speech/Music Classification Algorithm

Ⅰ. 서 론

최근 이동통신의 발전으로 무선통신기기를 이용한 멀티미디어 서비스가 보편화 되면서 제한된 주파수 대역의 효율적인 활용이 중요한 주제로 연구 되어 지고 있다. 제한된 통신망을 효율적으로 사용하기 위해 가변적인 전송률을 가지는 다양한 음성 코덱이 개발 되었는데[1~2], 음성신호의 유형에 따라 다른 전송률을 할당하기 위해 음성신호의 유형을 먼저 분별하는 작업이 필요하다. 이런 음성코덱 중의 하나인 ETSI의 3GPP2

* 정회원, ** 학생회원, 인하대학교 전자공학부
(Dep. of Electronics Engineering, Inha University)
*** 정회원, 한양대학교 융합전자공학부
(Dep. of Electronic Engineering, Hanyang University)
※ 이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (2009-0085162) 또한 본 연구는 지식경제부 및 한국산업기술평가관리원의 IT핵심기술개발사업의 일환으로 수행하였음. [KI001824, 장애인 및 고령자를 위한 Digital Guardian 기술개발]
접수일자: 2010년8월2일, 수정완료일: 2010년12월2일

selectable mode vocoder (SMV)의 음성신호 유형의 분류를 support vector machine (SVM)을 이용해 개선하는 방법[3]이 제안되었다. 본 논문에서는 SVM의 음성/음악 분류성능을 개선하는데 목적이 있고 이를 SMV 코덱의 음성신호 유형의 분류에 적용하여 보고자 한다.

SVM은 패턴인식에 우수함을 인정받아 많이 이용 되고 있는 machine learning 기법의 하나로써 데이터 마이닝 분야는 물론, 얼굴인식, 생체인식, 문자인식, 그리고 음성인식 등 다양한 분야에 응용되고 있다[4~5]. SVM에서 커널함수는 매우 중요한 역할을 담당한다. 첫째로 주어진 패턴이 선형적 분류가 가능하지 않은 경우가 많이 있는데 이를 해결하기 위하여 커널함수를 도입하여 패턴을 고차원 특징공간으로 사상시킨 후 패턴을 선형적으로 분류하게 된다. 둘째로 최적화된 패턴 인식률을 위해서는 해당 패턴에 가장 잘 맞는 커널함수를 선택하는 것이 필요하다[6]. 커널함수로는 주로 radial basis function (RBF)이나 polynomial 함수가 쓰이는데 본 논문에서는 RBF에 초점을 맞춘다.

커널함수를 정하는 것도 중요하지만 일단 커널함수를 정하게 되면 커널 파라미터를 최적화 하는 것도 중요하다[7]. 이는 커널 파라미터의 값에 따라 SVM의 패턴인식 능력이 크게 좌우되기 때문이다. RBF의 경우 커널 width를 파라미터로 조절할 수 있는데 이 값이 영이면 SVM이 분류초평면을 구하지 못하게 되는 underfitting이 발생하고, 무한대면 모든 특징벡터가 서포트 벡터 (support vector)가 되는 overfitting이 발생하게 된다. 따라서 generalization error를 최소화하고 학습시간을 단축시키는 최적화된 커널 파라미터를 구하는 것이 중요한데 이를 위해서 많은 최적화 알고리즘이 제안되었다[6, 8~9].

이 알고리즘들은 모두 SVM을 학습시키는데 쓰이는 방법들이다. 반면에 SVM의 학습과는 별도로 입력특징벡터를 가공하여 패턴 인식률을 높이는 방법이 최근에 제안되었다[10]. 이 알고리즘에서는 입력 특징벡터의 각 성분에 다른 가중치를 부과하여 SVM의 성능을 향상시킨다. 이런 알고리즘은 SVM의 학습 알고리즘과는 독립적으로 개발, 적용이 가능하여 기존의 SVM 성능을 더욱 향상시킬 수 있는 잠재력이 있다. 이와 같은 기존의 학습 알고리즘과 독립적인 알고리즘 개발을 위해, 본 논문에서는 SVM의 학습에 중요한 역할을 했던 커널함수 파라미터가 패턴 판별에 미치는 영향을 분석하고 이를 통해서 커널함수 파라미터가 SVM 판별을 미세 조정할 수 있다는 것을 밝힌다. 또한 입력 프레임 간의 상관관계를 밝히고 이것을 통해 커널함수 파라미터 조정의 지침을 구한다. 커널함수 파라미터로 SVM의 판별을 조정할 수 있는 것은 일반적인 결과이므로 제안하는 알고리즘은 입력신호 내에 강한 상관관계가 존재하는 모든 경우에 적용이 가능하다.

본 논문은 다음과 같이 구성된다. Ⅱ장에서는 SMV에 사용되는 특징벡터에 대해서 설명하고 Ⅲ장에서는 RBF 커널 width 파라미터가 SVM의 판별에 미치는 영향을 분석한다. Ⅳ장에서는 입력신호의 각 프레임간의 상관관계를 이용하여 커널 파라미터를 지도하는 방법에 대해 소개한다. Ⅴ장에서는 실험 설정과 실험결과를 보이고, Ⅵ장에서 본 논문을 끝맺는다.

Ⅱ. SMV의 개요 및 특징벡터 검출

SMV는 ETSI의 3GPP2 표준 코덱으로서 extended code excited linear prediction (ex-CELP) 기반의 압축방식을 사용하는데, 사람의 청각 특성에 최적화된 모델을 사용하여 음성을 저 전송률로 압축하는데 효율적이다[11~12]. 또한, 한정된 주파수 대역을 효율적으로 활용하기 위해 프레임 단위로 4가지의 가변 전송률을 제공하며 이동국과 기지국 사이의 통신망 채널에 따라 동적으로 변환되는 4가지 모드를 지원한다. 이러한 다양한 평균 전송률을 제공하기 때문에 시스템의 효율성과 음질간의 균형을 선택적으로 조절 할 수 있다.

SMV에서의 음악 분류 과정은 먼저 음성 검출기 (voice activity detection, VAD)에서 입력 신호가 음성과 묵음 또는 주변 잡음으로 나뉜 후 음성으로 판별된 경우에만 거치게 되며 음성/음악 분류에 사용되는 파라미터들은 다음과 같다.

1. 이동 평균 에너지 $\overline{E}$

$$E = 0.75 \cdot E + 0.25 \cdot E \quad (1)$$

E는 프레임 에너지 이다.

2. 잡음/묵음의 이동 평균 반사계수 $\overline{k_N}(i)$

$$\overline{k_N}(i) = 0.75 \cdot \overline{k_N}(i) + 0.25 \cdot k_1(i)$$
$$i = 1,...,10 \quad (2)$$

3. 부분적 잔류 에너지의 이동 평균 $\overline{E_N^{res}}$

$$\overline{E_N^{res}} = 0.9 \cdot \overline{E_N^{res}} + 0.1 \cdot E^{res} \quad (3)$$

$\overline{E_N^{res}}$는 $\overline{k_N}$에 따라서 값이 새로워진다.

4. 정규화 된 피치 상관도의 이동 평균 $\overline{corr_P}$

$$\overline{corr_P} = 0.8 \cdot \overline{corr_P} + 0.2 \cdot (\frac{1}{5} \cdot \sum_{i=1}^{5} corr_p^B(i)) \quad (4)$$

$corr^{B_P}(i)$는 이전 프레임의 피치 상관도이다.

5. 주기적 계수 $\overline{c_{pr}}$

$$\overline{c_{pr}} = \alpha \cdot \overline{c_{pr}} + (1-\alpha) \cdot c_{pr} \quad (5)$$

α는 c_{pr}에 따라 값을 바꿔주는 정해진 가중치이다.

6. 음악 연속 계수의 이동 평균 $\overline{c_M}$

$$\overline{c_M} = 0.9 \cdot \overline{c_M} + 0.1 \cdot c_M \quad (6)$$

SMV의 VAD에서는 식 (1)~(5)로부터 나온 결과를 정해진 문턱 값과 비교하여 음성의 유무를 판단하며 음악의 분류는 $\overline{c_{pr}} \geq 18$ 또는 $\overline{c_M} > 200$이면 음악으로 판단한다.

Ⅲ. RBF의 커널 width 파라미터가 SVM의 판별에 미치는 영향

이 장에서는 RBF를 커널함수로 사용하는 SVM이 RBF의 커널 width가 변화할 때 어떤 영향을 받는지 고찰한다. 입력벡터 $\boldsymbol{X}$가 선형으로 분류가 가능한 경우, SVM의 판별함수는 다음 식과 같다.

$$f(\boldsymbol{X}) = \sum_{i=1}^{M} \alpha_i^* z_i < \boldsymbol{X_i^*}, \boldsymbol{X} > + b^* \quad (7)$$

$\boldsymbol{X_i^*}$는 학습에 의해 구해진 M개의 서포트 벡터 (support vector) 중 i번째 벡터이다. 최적화 바이어스 (optimization bias) b^*와 α^*는 학습에 의해 구해지는 quadratic programming problem의 해이다. 입력벡터가 선형으로 분류가 불가능 한 경우 커널함수를 사용한 SVM의 판별함수는 다음과 같다.

$$f(\boldsymbol{X}) = \sum_{i=1}^{M} \alpha_i^* z_i K(\boldsymbol{X_i^*}, \boldsymbol{X}) + b^* \quad (8)$$

커널함수로 RBF를 사용한다면 $K(\boldsymbol{X_i^*}, \boldsymbol{X})$는 다음과 같다.

$$K(\boldsymbol{X_i^*}, \boldsymbol{X}) = \exp\left(-\frac{\| \boldsymbol{X_i^*} - \boldsymbol{X} \|^2}{\sigma^2}\right) \quad (9)$$

실제로 커널 width는 σ이지만 시뮬레이션에서 변화시키는 값은 $1/\sigma^2$이기 때문에 본 논문에서는 $1/\sigma^2$을 커널 width 파라미터라고 부르기로 한다. $1/\sigma^2$에 작은 양의 수 δ를 더해서 커널함수를 변화시킨다고 가정하면 변화된 커널함수는 다음과 같다.

$$K'(\boldsymbol{X_i^*}, \boldsymbol{X}) = \exp\left(-\left(\frac{1}{\sigma^2} + \delta\right) \| \boldsymbol{X_i^*} - \boldsymbol{X} \|^2\right) \quad (10)$$

이 식을 전개해서 다시 써보면 다음식이 된다.

$$K'(\boldsymbol{X_i^*}, \boldsymbol{X}) = \exp\left(-\frac{1}{\sigma^2} \| \boldsymbol{X_i^*} - \boldsymbol{X} \|^2\right) \cdot \exp\left(-\delta \| \boldsymbol{X_i^*} - \boldsymbol{X} \|^2\right) \quad (11)$$

이것은 본래의 $K(\boldsymbol{X_i^*}, \boldsymbol{X})$에 $\exp\left(-\delta \| \boldsymbol{X_i^*} - \boldsymbol{X} \|^2\right)$이 곱해진 형태이다. $\| \boldsymbol{X_i^*} - \boldsymbol{X} \|^2$은 양수이고 δ도 양의 수로 가정했으므로 $\exp\left(-\delta \| \boldsymbol{X_i^*} - \boldsymbol{X} \|^2\right)$은 1보다 작은 양수가 된다. 즉 $1/\sigma^2$에 양의 수 δ를 더하게 되면 원래 커널함수의 값이 줄어들게 된다. 반대로 양의 수 δ를 뺀다고 가정하면 본래의 $K(\boldsymbol{X_i^*}, \boldsymbol{X})$에 1보다 큰 양의 수 $\exp\left(\delta \| \boldsymbol{X_i^*} - \boldsymbol{X} \|^2\right)$이 곱해지게 되어 값이 커지게 된다.

여기서 주목할 점은 커지거나 작아지는 비율이 두 벡터간의 거리의 제곱에 따른다는 것이다. 즉 두 벡터간의 거리가 멀다면 커널함수의 값이 상대적으로 조금 변화하고 거리가 가깝다면 커널함수의 값이 상대적으로 많이 변화한다는 것이다. 커널함수의 값의 변화는 이렇게 수학적으로 예측할 수 있지만 판별함수 $f(\boldsymbol{X})$는 커널함수의 값 뿐 만이 아니라 $\alpha_i^* z_i$도 영향을 미치므로 실제로 변하는 값을 표 1에 나타내었다.

표 1에서 첫 번째 행은 커널 width 파라미터 $1/\sigma^2$에 더해진 δ 값을 나타낸다. 두 번째 행은 판별식의 값이 양의 수에서 음의 수로 바뀐 경우의 수와 더해주기 전 판별식의 값이 양인 경우의 수와의 백분율 값이다. 세 번째 행은 반대로 판별식의 값이 음의 수에서 양의 수로 바뀐 경우의 수와 더해주기 전 판별식의 값이 음인 경우의 수와의 백분

표 1. 커널 width 파라미터 변화에 따른 판별함수 $f(\boldsymbol{X})$의 +/- 부호의 변화

Table 1. Impact of kernel width parameter on the polarity of $f(\boldsymbol{X})$.

	0.03	0.06	0.09	-0.03	-0.06	-0.09
+ → −	23.4	46.7	55.8	0.17	0.29	0.23
− → +	0.18	0.37	0.56	7.44	35.9	62.1

율 값이다. 이 표는 V장에서 설명되어질 테스트 파일들을 바탕으로 구해진 평균값들을 보여주고 있다.

양의 수가 더해지면 판별식의 값이 양에서 음으로 바뀌는 경향이 두드러지고 음에서 양으로 변하는 경우는 1% 미만이다. 반대로 음의 수가 더해지면 판별식의 값이 음에서 양으로 변하는 경우가 많고 양에서 음으로 변하는 경우는 역시 1% 미만이다. 즉 양의 수를 더하면 한 클래스로의 판별이 늘어나고 음의 수를 더하면 다른 클래스로의 판별이 늘어난다. 또한 δ이 커질수록 판별식의 부호가 바뀌는 경우의 수가 늘어남을 볼 수 있다. 이 성질을 이용하면 커널 width 파라미터를 이용하여 판별식의 결과를 어느 정도 미세 조정할 수 있다.

그러나 SVM의 성능향상을 위해서 어떻게 판별식의 값을 조절해야 하는지 알아야 이 성질을 잘 활용할 수 있다. 이를 위해서 다음 장에서는 판별식 값을 조절하는 판단의 근거를 구하는 방법을 소개한다.

Ⅳ. 입력신호의 각 프레임간의 상관관계를 이용한 SVM 개선방법

실험에서 사용되어진 음성/음악 신호는 음성구간, 음악구간, 그리고 무음구간으로 구성되어 지는데, 각 구간은 어느 정도의 길이를 가지며 반복된다. 그러므로 각 구간은 실시간 분류를 위해 대체로 많은 수의 프레임으로 구성된다. 그러므로 주어진 구간 안에서는 강력한 상관성으로부터 다음 프레임이 현재 프레임과 같은 종류일 확률이 아주 높다고 할 수 있다. 예를 들어서 현재와 과거의 몇 입력 프레임이 음악구간을 반영한다고 가정하면 다음프레임도 음악 프레임일 확률이 아주 높다는 것이다. 실제로 확률을 구해본 결과 거의 100%의 확률로 바로 전 프레임까지 연속된 동일 클래스의 프레임의 개수에 관계없이 현재 프레임은 동일한 클래스에 속했다.

그러나 위의 경우는 실제 각 프레임의 클래스를 바탕으로 구해졌고 실제로 SVM이 패턴분류를 하는 경우,

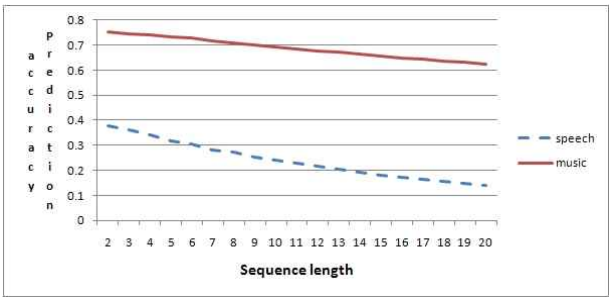


그림 1. SVM이 동일 클래스로 분류한 연속된 과거 프레임의 개수에 따른 현재 프레임이 같은 클래스일 확률

Fig. 1. Probability for the current frame to be the same class as previous frames with respect to the number of consecutive previous frames that has been classified identically.

이 정보는 존재하지 않는다. 프레임 간의 상관관계를 이용하여 각 프레임의 종류를 추정하는 데 쓸 수 있는 정보는 SVM의 패턴분류 결과뿐이다. 이 정보를 가지고 각 프레임의 클래스를 추정했을 경우, 실제 클래스와 비교하여 추정이 맞을 확률을 그림 1에 나타내었다.

SVM의 분류결과는 1 (음성)과 -1(음악)로 표현되어지고 점선은 음성을 나타내고 실선은 음악을 나타낸다. 가로축은 추정의 바탕이 되는 한 클래스에 속하는 연속된 프레임의 개수이고, 세로축은 현재 프레임이 과거 프레임과 동일하다고 추정한 경우의 정확도이다. 그림에 표시된 값들은 V장에서 설명될 테스트 파일에서 구해진 평균값이다.

과거 두개의 프레임이 동일한 클래스에 속한다고 SVM이 분류를 하였다면 현재 프레임도 같은 클래스에 속할 확률은 음악의 경우93%이고 음성의 경우 59%이다. 이것은 실제 프레임의 클래스를 사용했을 때 100% 가까운 확률에 비하면 음악의 경우는 다소 낮고 음성의 경우는 많이 낮은 확률이다. 한편 과거의 연속된 동일 클래스로 SVM에 의해 분류된 프레임의 개수가 늘어날수록 확률은 조금씩 늘어난다. 그러나 고려되는 과거 프레임의 개수가 늘어날수록 그 만큼 길게 연속되는 프레임의 빈도수는 확률의 증가분보다 더 많이 떨어지게 된다. 그러므로 과거 프레임들의 SVM에 의한 분류결과를 바탕으로 현재 프레임을 추정한다면 그림1처럼 추정의 정확도가 점차 떨어지게 된다.

그래서 과거 프레임들의 SVM분류결과를 가지고 추정하는 방법보다는 각 구간의 시작을 추정하는 방법을 채택하였다. 즉 예를 들어 과거의 연속된 프레임이 음성 클래스에 속한다면 그것은 음성 구간의 시작이라고

가정하여 다른 클래스의 시작이 추정될 때까지의 모든 프레임을 음성 프레임으로 간주한다. 이 기법은 각 프레임의 클래스를 추정하는 기법과 동일한 복잡도를 가지면서 더 높은 정확도를 보인다.

이런 프레임간의 상관관계를 이용하면 각 프레임의 클래스를 추정할 수 있다. 그리고 이 정보를 이용해서 3장에서 보인 RBF의 커널 width 파라미터를 조절하여 SVM의 판별결과를 미세 조정하는 기능을 잘 활용할 수 있다. 즉 음악일 가능성이 높은 프레임에는 음악 쪽으로 판별될 가능성을 높이는 방향으로 커널 width 파라미터를 조정해 주고, 음성일 가능성이 높은 프레임에는 음성으로 판별될 가능성으로 높이는 방향으로 파라미터를 조정해 준다.

V. 실 험

1. 실험 설정

제안된 알고리즘을 검증하기 위해서 제안된 알고리즘을 쓴 경우와 쓰지 않은 경우의 패턴 인식률을 비교하였다. 본 실험을 위해서 음성 데이터베이스로 8kHz로 샘플링 된 약 6 sec 정도의 깨끗한 음성으로 326명의 남자와 138명의 여자 화자에 의해서 화자마다 10개의 파일이 발음된 TIMIT 데이터베이스가 사용되었다[13]. 음악 데이터베이스는 CD로부터 다섯 가지 장르의 음악을 모바일 폰을 통해서 녹음하였고, 8kHz로 다운 샘플링 하여 사용하였으며, 각기 약 5분 정도의 길이를 가진다. 학습으로는 음성파일 4200개와 음악파일 50개(블루스 10개, 클래식 10개, 힙합 10개, 재즈 10개, 메탈 10개)가 사용되었다.

객관적인 평가를 위해 10-fold 교차검증을 수행하였으며 각 테스트 파일은 5개의 음성부분 (6~12초), 하나의 음악장르로 구성된 5개의 음악부분(28~32초), 10개 무음부분 (3~15초)으로 되어있다. 트레이닝 파일의 음악부분은 모든 장르의 음악이 혼합되었다. 성능 평가를 위해 테스트 파일의 20ms마다 실제 결과를 음성, 음악, 무음으로 분류하여 저장하고 SVM의 분류 결과와 비교하였다. 또한 제안된 알고리즘을 보다 공정하게 검증하기 위하여 음악과 음성이 무음구간 없이 자주 바뀌고 각 음악과 음성 부분의 길이가 짧은 특별한 테스트 파일들을 구성하였다. 각 테스트 파일에는 20개의 음성부분과 20개의 음악부분이 있고 각 부분은 5초의 길이를 가진다.

실험에 사용된 특징벡터로는 Ⅱ장에서 소개된 6가지의 파라미터를 벡터로 구성해 사용하였다.

SVM을 학습시키고 테스트 할 때 기본으로 사용되는 커널 width 파라미터 $1/\sigma^2$을 구하기 위해 여러 값을 시도하였고 그 중 가장 좋은 분류성능을 보이는 0.1이 사용되었다.

2. 실험 결과

Ⅳ장에서 설명하였듯이 현재 프레임의 추정을 바탕으로 커널 width 파라미터를 조절해 주어야 하는데 표 1에서 보인 것처럼 SVM의 결과가 양수에서 음수로 변화되려면 양의 수를 파라미터에 더해주어야 한다. 즉 음악은 SVM의 판별식 값이 음수여야 하므로 양수를 더해주고 반대로 음성은 음수를 더해주어야 한다.

그림 2는 커널 width 파라미터를 조절하면서 SVM의 분류 정확도가 어떻게 변하는지 알려준다. 이 그림은 metal 장르의 테스트 파일을 가지고 구한 결과이다. 다른 장르의 테스트 파일도 대부분 비슷한 결과를 보여주므로 본 논문에서는 소개하지 않는다. 가로축은 파라미터를 어떻게 조절하는지를 나타내고 세로축은 정확도를 나타낸다. 가로축에서 'p'는 현재 프레임의 추정이 음악일 때 더해주는 값을 나타내고 'n'은 현재 프레임이 음성으로 추정될 때 빼주는 값을 나타낸다. 예를 들어 'p0,02 n0.02'는 음악 프레임에는 0.02를 더해주고 음성 프레임에는 0.02를 빼준다는 의미이다. 커널 파라미터에 빼주는 값은 커널파라미터보다 작아야 한다. 커널파라미터보다 큰 경우에는 RBF의 특성상 함수값이 매우 커지기 때문에 음악/음성 분류성능이 오히려 떨어지게 된다. 즉 본 실험에서 사용된 0.1이 아니라 다른 커널파라

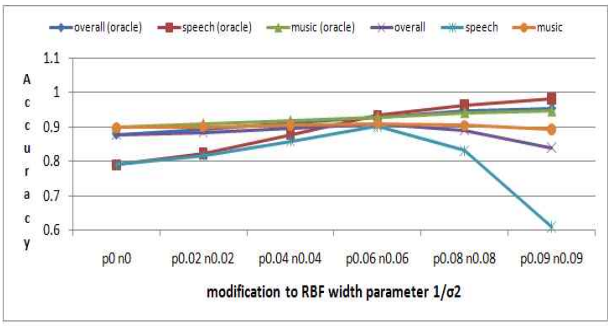


그림 2. RBF 커널 width 파라미터 $1/\sigma^2$의 조정에 따른 패턴분류 정확도의 변화

Fig. 2. Influence of RBF kernel width parameter $1/\sigma^2$ on the classification accuracy of the proposed SVM.

표 2. RBF 커널 width 파라미터 $1/\sigma^2$의 조정에 따른 패턴분류 변화에 대한 분석(프레임 클래스 추정에 실제 프레임 정보 사용한 경우)

Table 2. Detailed analysis on the changes in pattern classification made by adjusting RBF kernel width parameter $1/\sigma^2$ (actual frame information is used for predicting frame classes).

	p0 n0	p0.02 n0.02	p0.04 n0.04	p0.06 n0.06	p0.08 n0.08	p0.09 n0.09
S_{c-c}	14.79	14.78	14.78	14.78	14.79	14.79
M_{c-c}	73.11	73.06	72.99	72.96	72.91	72.91
S_{ic-c}	0	0.63	1.65	2.68	3.25	3.57
M_{ic-c}	0	0.86	1.66	2.56	3.7	4.1
S_{ic-ic}	3.9	3.27	2.25	1.23	0.65	0.34
M_{ic-ic}	8.2	7.34	6.54	5.64	4.5	4.1
S_{c-ic}	0	0.01	0.01	0.01	0	0
M_{c-ic}	0	0.05	0.12	0.15	0.21	0.21

표 3. RBF 커널 width 파라미터 $1/\sigma^2$의 조정에 따른 패턴분류 변화에 대한 분석(프레임 클래스 추정에 SVM 분류결과를 사용한 경우)

Table 3. Detailed analysis on the changes in pattern classification made by adjusting RBF kernel width parameter $1/\sigma^2$ (outpus of an SVM is used for predicting frame classes).

	p0 n0	p0.02 n0.02	p0.04 n0.04	p0.06 n0.06	p0.08 n0.08	p0.09 n0.09
S_{c-c}	14.79	14.68	14.58	14.47	13.27	9.62
M_{c-c}	73.11	72.72	72.39	72.05	70.79	69.33
S_{ic-c}	0	0.56	1.49	2.4	2.28	1.8
M_{ic-c}	0	0.54	1.26	1.94	2.76	3.27
S_{ic-ic}	3.9	3.34	2.42	1.51	1.63	2.1
M_{ic-ic}	8.2	7.65	6.94	6.26	5.43	4.92
S_{c-ic}	0	0.11	0.21	0.31	1.52	5.17
M_{c-ic}	0	0.39	0.73	1.06	2.32	3.78

미터 값을 사용한 경우, 그 값을 고려하여 파라미터에 더하거나 빼는 값을 정해야 한다.

그림에는 6개의 선이 있는데 이는 음성, 음악, 그리고 전체 정확도를 나타낸다. 'oracle'이라고 되어 있는 선들은 실제 프레임의 클래스를 가지고 파라미터를 조절해 준 결과이고 그렇지 않은 선들은 SVM의 판별결과를 바탕으로 파라미터를 조절해 준 결과이다.

실제 프레임의 클래스를 바탕으로 한 경우는 파라미터의 변화가 클수록 표 1에서처럼 더 많은 수의 잘못된 분류가 고쳐져서 정확도가 높아진다. 그러나 SVM의 분류결과를 바탕으로 한 경우, 'p 0.08 n 0.08' 이상에서는 오히려 정확도가 떨어지는 것을 볼 수 있다. 그 이유를 알아보기 위해 표 2 와 표 3에서 구체적으로 패턴분류에 어떤 변화가 있었는지 나타내었다. 이 표들도 그림 2와 동일하게 metal 장르의 테스트 파일을 바탕으로 작성되었다.

표 2는 현재 프레임 클래스 추정에 실제 클래스를 사용한 경우이고 표 3은 SVM의 분류결과를 사용한 결과이다. 맨 왼쪽 열은 커널 width 파라미터를 조정해 생기는 변화를 세부적으로 나누어 놓은 것이다. 'S'는 음성 (speech)를 의미하고 'M'은 음악 (music)을 뜻한다. 그리고 아래첨자 'c'는 분류가 맞는다(correct)는 의미이고 'ic'는 틀린다(incorrect)를 뜻한다. 예를 들어 'S_{ic-c}'는 음성 프레임이 파라미터 조정 전에는 음악 프레임으로 잘못 분류되었다가 조정 후에는 음성으로 맞게 분류된 경우를 뜻한다. 'M_{ic-ic}'는 조정 전이나 조정 후 모두 음성 프레임으로 잘못 분류된 음악 프레임을 뜻한다. 표의 숫자들은 각 해당 경우의 전체 프레임 개수에 대한 백분율이다.

표 2를 보면 잘못 분류되었다가 파라미터 조정 후에 맞게 분류된 경우 (S_{ic-c}와 M_{ic-c})의 비율이 늘어난 것을 볼 수 있고 따라서 파라미터 조정 전과 후 모두 틀리게 분류된 비율 (S_{ic-ic}와 M_{ic-ic})이 줄어든 것을 알 수 있다. 그러므로 그림 2에서의 'oracle'의 이상적인 결과를 설명해준다. 그리고 표 3을 보면 그림 2에서 왜 향상되던 정확도가 중간에 떨어지는 지 알 수 있다. 파라미터가 증가하고 프레임의 추정이 완벽하지 않으므로 더 많은 수의 조정 전에는 맞던 분류가 조정 후에 틀리게 바뀌었고 (S_{c-ic}와 M_{c-ic}), 파라미터 조정에 상관없이 맞던 분류 (S_{c-c}와 M_{c-c})도 줄어들게 되었다. 또한 이것으로 인해 프레임 추정의 정확도가 더 떨어지게 되어 파라미터 조정으로 맞게 분류가 되던 경우 (S_{ic-c})의 비율이 줄어들고 그래서 파라미터조정으로도 분류가 계속 틀린 경우(S_{ic-ic})는 늘어나게 되었다.

마지막으로 제안된 기법으로 인한 성능개선을 알아보도록 한다. 표 4는 제안된 기법을 적용했을 때와 적용하지 않았을 때의 성능을 비교함으로써 제안된 기법의 가능성을 알 수 있다.

표의 왼쪽에서 첫 번째 열에 있는 'special'은 공정한 검증을 위해 특별히 다수의 짧은 길이를 가진 음성과 음악 부분으로 구성된 테스트 파일에 대한 결과이다. 표의 왼쪽에서 두 번째 열은 사용된 기법을 나타내는데 'SVM'은 제안된 기법이 사용되지 않은 경우이고, 'eSVM'은 'enhanced SVM'의 뜻으로 SVM의 분류결과로 현재 프레임을 추정하는 제안된 기법이 적용된 경우를 뜻한다. 'eSVMo'는 'enhanced SVM with oracle

표 4. 제안된 기법을 적용했을 때와 안했을 때의 음성/음악 분류 성능 비교

Table 4. Comparison of speech/music detection probability Pd and total error probability Pe between an SVM with the proposed enhancement and an SVM without the proposed enhancement.

Class	Method	Music P_d	Speech P_d	Total P_e
Blues	SVM	0.87	0.85	0.13
	eSVMo	0.93	0.97	0.06
	eSVM	0.89	0.94	0.1
Classic	SVM	0.66	0.74	0.33
	eSVMo	0.84	0.89	0.15
	eSVM	0.69	0.81	0.29
Hiphop	SVM	0.91	0.75	0.12
	eSVMo	0.95	0.94	0.05
	eSVM	0.94	0.85	0.08
Jazz	SVM	0.91	0.75	0.12
	eSVMo	0.95	0.94	0.05
	eSVM	0.94	0.85	0.08
Metal	SVM	0.85	0.76	0.17
	eSVMo	0.92	0.95	0.07
	eSVM	0.87	0.85	0.14
Special	SVM	0.82	0.87	0.16
	eSVMo	0.91	0.96	0.06
	eSVM	0.86	0.91	0.12
Avg	SVM	0.84	0.79	0.17
	eSVMo	0.92	0.94	0.07
	eSVM	0.87	0.87	0.14

frame information'의 뜻으로 실제 프레임 정보를 바탕으로 하는 제안된 기법을 말한다. P_d는 각 음성과 음악이 정확하게 분류될 확률이고 P_e는 $(1-P_d)$로서 음성과 음악을 합친 error probability이다. 실험 시 그림 2에 나타난 것처럼 커널 파라미터에 더해지는 수를 여러 가지로 변화시켰는데, 표 4에 표시된 결과를 구할 때는 가장 낮은 Pe를 가지면서 음성, 음악 모두 성능이 향상되는 한 경우를 선택하였다.

'eSVMo'와 'eSVM'의 차이는 프레임의 클래스 추정 방법인데, 이것이 큰 차이를 만들고 많은 개선의 여지가 남아있는 것으로 보인다. 'eSVMo'는 현실적으로는 구현 불가능하므로 'eSVM'의 개선여지를 파악에 도움을 주는 역할이다. 현실적으로 구현 가능한 'eSVM'은 모든 장르의 테스트뿐만 아니라 특별히 구성된 테스트에서도 고르게 성능을 향상시키고 기법을 안 쓴 경우보다 음성의 경우 8%, 음악의 경우 3%, 전체적으로 3% 정도 성능을 향상 시킨다.

제안된 기법과 동일하게 클래스 분류 시 적용되는 다른 기법[10]과 비교해 보았을 때 본 기법은 음성분류 성능 향상에 강점을 보이고 기존의 기법은 음악분류 성능 향상에 강점을 보인다. 그러나 주목해야할 점은 두 기법은 함께 적용이 가능하다는 것이다. 제안된 기법은 또한 학습 시에 사용되는 모든 기법과도 병행이 가능한 강점을 가지고 있다.

VI. 결 론

본 논문에서는 SVM의 음악/음성 분류성능을 향상시키기 위해 패턴 판별 시에 RBF 커널함수의 width 파라미터를 입력 프레임 간의 상관관계를 이용한 프레임의 클래스 추정을 바탕으로 미세 조정하는 방법을 제안 하였다. 그리고 ETSI의 3GPP2 표준코덱인 SMV의 실시간 음성/음악 분류에 적용하여 보았다. 이 기법은 SVM의 성능을 향상시킬 뿐 아니라 다른 기법들과도 병용할 수 있다는 장점도 가지고 있다. 실험을 통하여 검증한 결과, SMV의 음성/음악 분류 성능 향상에의 충분한 가능성을 확인할 수 있었다.

앞으로의 연구과제로는 RBF의 커널 width 파라미터와 SVM 판별 값의 보다 정확한 상관관계를 구하여 파라미터의 조정을 통한 SVM 판별의 조절능력을 향상시키는 것과 현재 프레임의 클래스를 간단하면서 보다 정확하게 추정하는 것 등이 있다. 또한 라디오방송을 녹음하여 실험데이터로 사용하여 제안된 기법의 실제적 응용 가능성을 가늠해 볼 계획이다.

감사의 글

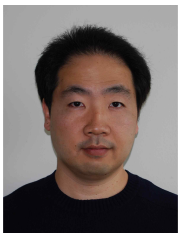
이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (2009-0085162) 또한 본 연구는 지식경제부 및 한국산업기술평가관리원의 IT핵심기술개발사업의 일환으로 수행하였음. [KI001824, 장애인 및 고령자를 위한 Digital Guardian 기술개발]

참 고 문 헌

[1] 3GPP2 Spec., "Source-controlled variable-rate multimedia wideband speech codec (VMR-WB), service option 62 and 63 for spread spectrum systems," *3GPP2-C.S0052-A*, vol. 1.0, April. 2005.

[2] Y. Gao, E. Shlomot, A. Benyassine, J. Hyssen, Huan-yu Su, and C. Murgia, "The SMV algorithm selected by TIA and 3GPP2 for CDMA appications," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 709-712, May 2001.
[3] S. -K. Kim and J. -H. Chang, "Speech/music classification enhancement for 3GPP2 SMV codec based on support vector machine," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, Vol. E92-A, no. 2, February 2009.
[4] X. Wang, J. Chen, P Wang, Z. Huang, "Infrared human face auto locating based on SVM and a smart thermal biometrics system," in *Proc. Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06)*, vol. 2, pp. 1066-1072, October 2006.
[5] A. Ganapathiraju, J. E. Hamaker, J. Picone, "Applications of support vector machines to speech recognition," *IEEE Trans. Signal Processing*, vol. 52, pp. 2348-2355, August 2004.
[6] L. -P. Bi, H. Huang, Z. -Y. Zheng, and H. -T. Song, "New heuristic for determination Gaussian kernel's parameter," in *Proc. International Conference on Machine Learning and Cybernetics*, vol. 7, pp. 4299-4304, August 2005.
[7] S. S. Keerthi, C. -J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," Neural Computation, vol. 15, pp. 1667-1689, 2003.
[8] J. Tian and L. Zhao, "Weighted Gaussian kernel with multiple widths and network kernel pattern," in *Proc. International Symposium on Information Engineering and Electronic Commerce*, pp. 379-382, May 2009.
[9] N. E. Ayat, M. Cheriet, and C. Y. Suen, "Automatic model selection for the optimization of SVM kernel," *Pattern Recognition*, vol. 38, pp. 1733-1745, October 2005.
[10] S. -K. Kim and J. -H. Chang, "Discriminative weight training for support vector machine-based speech/music classification in 3GPP2 SMV codec," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. E93-A, no. 1, pp. 316-319, January 2010.
[11] S. C. Greer, and A. Dejaco, "Standardization of the selectable mode vocoder," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 953-956, May 2001.
[12] C. V. Goudar, P. Rabha, M. Deshpande, and A. Rao, "SMVLite: reduced complexity selectable mode vocoder," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 701-704, May 2006.
[13] W. M. Fisher, G. R. Doddington and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop Speech Recognition*, pp. 93-99, February 1986.

저 자 소 개



임 정 수(정회원)
1996년 인하대학교 전기공학과 학사.
2004년 University of Maryland ECE 석사.
2009년 North Carolina State University ECE 박사.
2010년 인하대학교 박사후연구원
<주관심분야 : 컴퓨터 구조, 임베디드 시스템, 신호처리>

송 지 현(학생회원)
2007년 인하대학교 전자전기공학부 학사.
2009년 인하대학교 전자공학과 석사
2010년 인하대학교 전자공학과 박사과정.
<주관심분야 : 디지털신호처리>

장 준 혁(정회원)
1998년 경북대하교 전자공학과 학사.
2000년 서울대학교 전기공학부 석사.
2004년 서울대학교 전기컴퓨터공학부 박사.
2000년～2005년 (주)넷더스 연구소장
2004년～2005년 캘리포니아 주립대학, 산타바바라(UCSB) 박사후연구원
2005년 한국과학기술연구원(KIST) 연구원
2005년～2011년 인하대학교 전자공학부 조교수
2011년 현재 한양대학교 융합전자공학부 부교수
<주관심분야 : 음성신호처리, 오디오 신호처리, 통신 신호처리, 휴먼/컴퓨터 인터페이스>