# Support Vector Machines에 의한 음소 분할 및 인식

이광석* · 김현덕*

*진주산업대학교

# Phoneme segmentation and Recognition using Support Vector Machines

Gwang-seok Lee* · Deok-hyun Kim*

*Jinju National University

E-mail : kslee@jinju.ac.kr

## 요 약

우리는 본 연구에서 학습방법으로서 연속음성을 초성, 중성, 종성의 음소단위로 분할하기 위하여 인공 신경회로망의 하나인 SVMs을 사용하였으며 분할한 음소단위의 음성으로 연속음성인식에 적용하여 그 성능을 살펴보았다. 음소경계는 단 구간에서의 최대 주파수를 가진 알고리듬에 의하여 결정되며 또한 음성인식처리는 CHMM에 의하여 이루어지며 목측에 의한 분할결과와도 비교하여 살펴보았다. 시뮬레이션 결과로부터 초성의 분할성능에서 제안한 SVMs를 적용한 결과가 GMMs보다 효율적인을 알 수 있었다.

## ABSTRACT

In this paper, we used Support Vector Machines(SVMs) as the learning method, one of Artificial Neural Network, to segregated from the continuous speech into phonemes, an initial, medial, and final sound, and then, performed continuous speech recognition from it. A Decision boundary of phoneme is determined by algorithm with maximum frequency in a short interval. Speech recognition process is performed by Continuous Hidden Markov Model(CHMM), and we compared it with another phoneme segregated from the eye-measurement. From the simulation results, we confirmed that the method, SVMs, we proposed is more effective in an initial sound than Gaussian Mixture Models(GMMs).

## 키워드

Phoneme Segmentation, Pattern Recognition Support Vector Machines, Continuous Hidden Markov Model

## Ⅰ. 서 론

음성에 대한 연구가 발전함에 따라 음성신호의 분할과 라벨링된 음성 DB에 대한 필요가 증가되고 있으며 HTK와 같은 음성인식도구에서는 발음사전을 바탕으로 자동으로 음성인식의 단위를 정렬하여 초기 음소모델을 구성하고 다이폰이나 트라이폰 모델 등으로 확장하여 사용하기도 하지만 안정된 발음사전의 구성과 데이터의 확보에 어려움이 있다. 반면, 음성을 일정 세그먼트의 연결로 가정하고 인식단위로의 자동 분할을 통한 음성인식시스템의 경우, 최소 인식단위인 음소 단위로 정확하게 분할되고 라벨링된 음성 DB는 인식기의 성능에 결정적인 영향을 미치게 된다.

현재로서는 음소 단위의 음성인식이 음성의 모델 수가 적어 많은 이점이 있는데도 불구하고 이러한 음성 DB를 만들기 위해선 많은 시간과 노력을 필요로 하며 이러한 분할구간의 자동결정은 한국어의 조음결합 현상 등을 고려할 경우 여전히 어려운 과제로 남아 있다. 또한 자음의 인식률 저하도 실제 적용상의 문제점으로 지적된다. 음성정보의 의미 단위로 자동으로 분할하기 위한 방법은 다양한 방법이 제안되어 있으나 본 연구에서는 최근 많은 연구가 활발히 이루어지고 있는 SVM분류기와 전통적인 패턴인식 방법인 GMM을 이용하여 자동으로 음성을 음소단위로 분할하였다. 실험결과, 초성의 경우 GMM보다 SVM이 높은 인식률을 보였고, 중성, 종성의 경우는 GMM이 조금 더 높은 인식률을 보임을 알 수 있었다.

본 연구의 구성은 2장에서 선형공간과 비선형공간에서의 SVMs 분류기에 대해서 간단하게 설명하였고, 3장에서는 음성의 한 프레임마다 SVMs로 분류된 음성을 자동 음절분할하기 위한

후처리 과정을 설명하였다. 실험에 사용된 DB 및 실험 결과를 4장에서 기술하였으며, 마지막으로 5장에서 결론 및 향후 과제를 제시하였다.

## Ⅱ. Support Vector Machines(SVMs)

### 2.1 선형 SVMs

$y_i = \{-1,+1\}$인 두개의 클래스에 속하도록 $(x_1,y_1),\cdots,(x_l,y_l),\in R^N$ 학습 벡터를 분류한다고 고려하자. 여기서 우리는 초평면을 이용해서 두 개의 클래스로 분류할 것이다.

$$(w\cdot x)+b=0,\quad w\in R^N \text{and} b\in R \tag{1}$$

여기서, *w*와 *b*는 *f(x)=sign(w·x+b)*인 판별 함수에서 적용될 파라미터들이다.

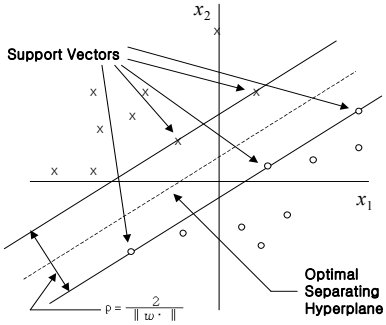


그림 1. 선형 공간에서의 분류

그림 1에서와 같이 2차원 공간의 경우 입력 데이터들을 분류할 수 있는 선형 분류기가 다수가 존재할 수 있다. 그러나 초평면과 여기에 가장 가까운 데이터들의 거리를 최대화하는 초평면은 단 하나만 존재한다. 이러한 선형 분류기를 Optimal Separating Hyper plane(OSH)라 부르며 초평면은 $(w\cdot x)+b=0$은 다음조건을 만족한다.

$$\begin{aligned}(w\cdot x)+b>0,\quad &\text{if } y_i=1\\ (w\cdot x)+b<0,\quad &\text{if } y_i=-1\end{aligned} \tag{2}$$

식(2)의 *w*와 *b*를 적절하게 선택함으로써 다음의 제약식(3)을 만족하는 하나의 분류평면으로 수식화할 수 있다.

$$y_i[(w\cdot x)+b]\geq 1,\ i=1,\ \cdots,\ l \tag{3}$$

그리고 초평면을 최적으로 만드는 Cost Function은 다음의 (4)식과 같다.

$$\Phi(w)=\frac{\|w\|^2}{2} \tag{4}$$

따라서 최적화문제는 Lagrangian 곱셈기를 이용하여 등가 비제약 최적화문제로 다시 정의되고 Lagrange함수에 의해 구함으로써 해결된다.

$$L(w,b,a)=\frac{\|w\|^2}{2}-\sum_{i=1}^{l}a_i\{[(w\cdot x)+b]y_i-1\} \tag{5}$$

식 (5)를 KKT조건에 의해 정리하면 다음과 같다.

$$W(a)=\sum_{i=1}^{1}a_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_ia_jy_iy_j(x_i\cdot x_j) \tag{6}$$

식(6)의 해를 찾기 위해서는 2차 계획법(Quadratic Programming)과 대수적 방법을 필요로 하며 OSH는 다음에 의해 구해진다.

$$W_0=\sum_{i=1}^{1}a_{o,i}d_ix_i,b_o=-\frac{1}{2}W_o[x_r+x_s] \tag{7}$$

여기서, $X_r$과 $X_s$는 각 클래스의 Support Vector들이고 그림 1에서 초평면에 각각 가장 가까운 점들이다.

### 2.2 비선형 SVMs

비선형 데이터 공간에서는 제약식을 위반하는 양을 평가하는 새로운 변수 $\xi_i$를 이용함으로써 마진을 최대화한다.

$$\begin{aligned}&\min\Phi(W)=\frac{\|w\|^2}{2}+C(\sum\xi_i)\ \ y_i[(w\cdot x+b)]\geqq 1-\xi_i\\ &\textit{Subject to} \text{ and } \xi_i\geqq 0 i=1,\ldots,l\end{aligned} \tag{8}$$

여기서 다시 재 정의된 Lagrange함수는 다음과 같다.

$$\begin{aligned}L(w,b,a)=&\frac{\|w\|^2}{2}+C\left[\sum_{i=1}^{l}\right]-\sum_{i=1}^{l}r_i\xi_i\\ &-\sum_{i=1}^{l}a_i\{[(w\cdot x)+b]y_i-1+\xi_i\}\end{aligned} \tag{9}$$

여기서, $r_i$과 $\xi_i$는 식 (8)의 제약식과 관련되며 $a_i$의 값은 $0\leqq a_i\leqq C$ 를 만족해야 한다. 만약 선형 경계가 부적절하고 비선형 분류평면일 경우 입력 벡터 *x*를 고차원의 특징 공간으로 Mapping할 수 있으며 그 역할은 *K(x,y)*가 담당한다.

$$W(a)=\sum_{i=1}^{1}a_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_ia_jy_iy_jK(x_i\cdot x_i) \tag{10}$$

2.3 다중 클래스 분류 기법

SVMs의 이전 분류능력을 다중으로 확장하기 위해서 OPC(One-Per-Class), PWC(paire Wise Coupling)이 일반적으로 알려져 있으며 non-sense output 때문에 이를 보완하기 위한 다양한 방법들이 모색 중에 있으며 본 연구에서는 OPC를 적용하였다.

## III. 자동 음소분할 처리

SVM을 이용한 멀티 클래스 분류를 통하여 초성·중성·종성에 대한 프레임 단위의 1차 분류를 기반으로 하여 2차 음소경계를 결정하기 위해서는 1차 분류 결과의 오인식에 대한 보상처리가 필요하다. 이를 위해서 본 연구에서는 1차 인식결과에 단구간의 최빈값을 이용하는 후처리 알고리즘을 제안하고 적용하였다.

*Step 1* : 멀티 클래스 분류 결과를 초·중·종성으로 Mapping처리
*Step 2* : 시작 프레임과 종료 프레임의 전, 후에 의사 보상 값을 패딩, 패딩 프레임의 수는 짝수
*Step 3* : 단구간 프레임 크기를 한 프레임씩 슬라이딩하면서 최빈 인덱스를 보상한 후 인덱스로 설정
*Step 4* : (패딩 프레임 수/2)의 프레임을 전후 제기
*Step 5* : Step 4의 결과를 바탕으로 최종 음소 경계 결정

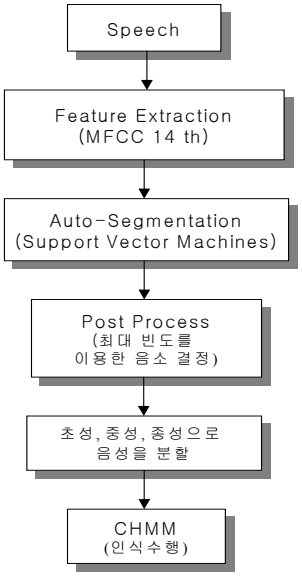


그림 2. 자동 음소분할 처리 흐름도

## IV. 실험결과 및 고찰

4.1 시뮬레이션 조건

음소분할 실험을 위하여 CVC형 108음절 DB를 구성하였다. 본 DB는 우리말 음성의 CVC형 음절로 초성(ㅂ, ㄷ,ㄱ, ㅍ, ㅌ, ㅋ), 중성(ㅏ, ㅓ, ㅗ, ㅟ, ㅜ, ㅣ), 종성(ㄴ, ㄹ, ㅁ)으로 이루어진 120개의 유사 음절로 구성되어져 있으며 5명의 화자가 5회 발성하여 3회분은 학습용으로 나머지 2회분은 평가용으로 이용하였다. 목측에 의하여 초성, 중성, 종성을 분리한 음소 DB를 별도 구성하여 학습을 행하고 분할과 인식실험을 행하였으며 음성의 분석조건은 표 1과 같다.

표 1. 음성분석조건

| | |
|---|---|
| A/D | 16kHz, 16bit |
| Filtering | LPF, 7KHz |
| Step Size | 60 point |
| Window Length | 256 point |
| Feature Parameter | MFCC 14th |

4.2 시뮬레이션 방법 및 결과

분할 방법은 SVM과 GMM으로 각각 행하고 그 결과를 서로 비교하였다. SVM은 초·중·종성으로 3개의 클래스로 하였고 GMM은 초·중·종성의 15개 클래스로 다중 분류를 행하고, 이를 초·중·종성으로 매핑하여 제안하는 후처리 알고리듬으로 그 결과를 보상하는 방법으로 음소 경계를 결정하였다.

표 2. 자동분할의 성능(목측분할과의 편차)

| Frame | | 초성 | 중성 | 종성 |
|---|---|---|---|---|
| GMM | 평균 | 6.32 | 4.94 | 4.68 |
| | 표준편차 | 6.86 | 6.24 | 6.35 |
| SVM | 평균 | 3.42 | 7.16 | 6.58 |
| | 표준편차 | 3.52 | 8.30 | 8.90 |

표 3. 음소단위 분할 후의 인식성능(%)

| Error | GMM분할 | SVM분할 | 목측분할 |
|---|---|---|---|
| 초성 | 23.80 | 20.93 | 13.89 |
| 중성 | 5.46 | 7.79 | 3.98 |
| 종성 | 8.15 | 9.07 | 5.37 |

표2와 표3에서 보듯이 시뮬레이션 결과로 알 수 있듯이 초성에서는 SVM이 중성, 종성에서는 GMM이 성능이 비교적 우수함을 확인할 수 있었다.

## Ⅴ. 결 론

우리말은 외국어와 달리 초성, 중성, 종성이 합해져서 음질을 이루고 이 음절이 단어와 문장을 이루기 때문에 인식단위, 특히 음소와 같은 최소 단위로의 안정된 분할은 연속음성인식을 위한 주목할 만한 연구과제이다. 본 연구에서는 SVM을 이용해서 자동 음소분할을 시도하고 이를 통하여 음성인식을 행하고 GMM을 이용한 것과 그 성능을 서로 비교하였다.

시뮬레이션 결과 음성의 SVM은 초성에서 GMM은 중성, 종성에서 각각 비교적 우수한 분할 성능을 확인했다. 현재 SVM은 많은 연산량과 학습시간을 필요로 하지만 프로세서의 성능의 비약적인 발전으로 문제가 되지 않으며 이는 무성음, 파열음, 마찰음 등의 비선형성을 가지는 음소 모델에선 확률 및 통계보다 나은 접근 방법으로 확인되고 있다. 따라서 SVM을 보다 더 최적화하기 위한 여러 가지 방법들이 연구되어야 할 것으로 생각되며 향후 SVM을 통한 최적의 음소 인식기를 구현하여 연속 음성인식에 적용할 계획으로 연구를 계속 진행할 것이다.

## 참고문헌

[1] Bernhard Scholkopf, Alexander J. Smola, "Learning with Kernels", The MIT Press, 2002

[2] Christopher J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Bell Laboratories, Lucent Technologies. 2008.

[3] J. Weston, C. Watkins, "Multi-class Support Vector Machines, Technical Report, Royal Holloway, University of London. 2008.

[4] Yonas B. Dibike, Slavco Velickov, Dimitri Solomatine, Michael B. Abbott, "Model Introduction with Support Vector Machines: Introduction and Applications", ASCE Journal of Computing in Civil Engineering, Vol.15, No.3, pp.208-216, 2007.

[5] Edgar E. Osuna, Robert Freund and Federico Girosi, "Support Vector Machines: Training and Applications", C.B.C.L Paper No.144, 2007.

[6] Philip Clarkson and Pedro J. Moreno, "On the Use of Support Vector Machines for Phonetic Classification", ICASSP 2005.

[7]fttp://fttp.idiap.ch/pub/learning/SVMTorch.tgz