논문 21-46-10-23 The Journal of Korean Institute of Communications and Information Sciences '21-10 Vol.46 No.10
https://doi.org/10.7840/kics.2021.46.10.1741

SVM을 이용한 음성신호 감정분류에 관한 연구

염 정 석•, 유 광 복°, 장 경 남*

A Study on the Emotion Classification of the Speech Signal Using Support Vector Machine

Jeong-seok Yeom•, Kwang-Bock You°, Kyungnam Jang*

요 약

본 논문은 제약된 환경에서 음성신호의 감정을 분류하는 알고리즘을 제안한다. 제안한 알고리즘은 SVM을 분류기로 이용하고 Autocorrelation Function (ACF)는 피치 주기를 구하는데 사용하였으며 Linear Predictive Coding (LPC)와 Split-LPC으로는 포먼트 주파수를 측정하였고, Zero Crossing Rate (ZCR)와 Short Time Energy (STE)는 유성음과 무성음을 판별하는 알고리즘의 개발에 활용되었다. 남성 3명, 여성 2명의 음성 데이터로 실험을 진행하였다. 본 논문에서는 SVM을 사용하여 감정을 분류하기 전 성별 분류의 과정을 추가하는 방식을 제안하여 정확도를 향상시켰다. 성별 및 감정분류는 피치 주기와 다수의 포먼트 주파수 사이의 관계를 그래프로 나타내어 확인하였다. 제안한 알고리즘은 제약된 환경에서도 향상된 정확도로 감정분류를 하였고 이에 화자의 판별된 감정에 대응하는 반응을 제공할 수 있는 가능성을 보였다.

키워드 : 감정분류, Autocorrelation Function (ACF), Support Vector Machine (SVM), Linear Predictive Coding (LPC), Split-LPC

Key Words : Emotion classification, Autocorrelation Function (ACF), Support Vector Machine (SVM), Linear Predictive Coding (LPC), Split-LPC

ABSTRACT

This paper proposes an algorithm to classify the emotions of speech signals in constrained environments. The proposed algorithm used the SVM as a classifier, the Autocorrelation Function (ACF) was used to obtain pitch period, the Linear Predictive Coding (LPC) and Split-LPC were used to measure formant frequencies, and Zero Crossing Rate (ZCR) and Short Time Energy (STE) were used to determine voiced and unvoiced sounds. The experiment was conducted with voice data of three men and two women. In this paper, propose a method to add a process of gender classification before classifying emotions using SVMs to improve accuracy. Gender and emotional classification were graphically identified the relationship between pitch period and multiple formant frequencies. The proposed algorithm made improved accuracy emotion classification in constrained environments, and showed the potential to provide responses to the speaker's discriminated emotions.

※ 본 연구는 2020년도 숭실대학교 교내연구비 지원에 의한 연구임
• First Author : Soongsil University School of Electronic Engineering, jswizard@naver.com, 학생회원
° Corresponding Author : Soongsil University School of Electronic Engineering, kwangbockyou@ssu.ac.kr, 정회원
* Soongsil University, Department of Korean Language & Literature, knjang@ssu.ac.kr
논문번호 : 202108-190-0-SE, Received July 29, 2021; Revised August 24, 2021; Accepted August 24, 2021

Ⅰ. 서 론

AI (Artificial Intelligence) 스피커 또는 AI 비서가 많은 IT 기업들에서 개발, 상용화 중이다. AI 스피커의 가장 큰 특징인 '음성'은 컴퓨터와 사람의 상호작용에 있어 기존의 입력 방식에 비해 잠재력이 있다. 이러한 특징으로 AI 스피커는 계속적인 시장성장 중이며 이상적인 차세대 인터페이스로 주목받고 있다. 현재까지의 AI 제품들은 사용자의 음성을 분석해 사용자의 감정에 따른 변화된 서비스를 제공하지는 않는다. 감정변화에 따른 적절한 서비스를 제공한다면 다양한 측면에서 사용자 만족도가 상승할 것으로 예상된다.

본 논문은 빅데이터를 얻을 수 없는 제한된 환경에서 음성신호를 기반으로 인간의 기본적인 감정을 분류할 수 있는 알고리즘을 제안 하였다. 이에 제한된 환경에서도 AI 제품의 사용자에게도 감정에 대응하는 다양한 서비스를 제공하는 것을 가능할 수도 있을 것이다.[1]

SVM 분류기의 파라미터로 음성신호의 피치 주기와 포먼트 주파수들을 활용하였다. 먼저, 음성신호에서 유성음과 무성음을 분리하기 위해서 ZCR과 STE를 활용한 간단한 알고리즘을 개발하였다. 유성음에서 피치 주기를 측정할 수 있는 시간 도메인의 두 함수 ACF와 AMDF를 설명하였고 본 논문에서는 ACF를 사용하였다. 그리고 포먼트 주파수들을 추정하기 위해서 LPC 분석을 설명하였다. 특별히 이런 LPC 분석에 나타나는 불확실성을 제거하고 좀 더 선명한 포먼트 주파수 값을 얻기 위해서 10차의 합성 필터를 5개의 2차의 필터로 분리하여서 직렬로 연결하여 분석하는 Split-LPC 알고리즘을 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 피치 검출 함수인 ACF 함수와 AMDF 함수에 대하여 간략히 기술하고, 3장에서는 LPC 분석과 Split-LPC 분석의 차이를 보였다. 4장에서는 유무성음 판별을 제공하는 알고리즘을 설명하였다. 이어서 5장에서는 SVM 파라미터들을 얻는 실험을 수행하였고, 6장에서 SVM 분류기를 적용하는 알고리즘을 제안하고 이를 Matlab을 이용하여 시뮬레이션 결과를 보였다. 마지막으로 7장에서 본 논문의 결론을 설명하였다.

Ⅱ. Pitch Detection Algorithms

2.1 ACF

좌우 대칭의 우함수 (even function)인 ACF는 아래의 (1)식으로 정의된다. 특별히 ACF는 lag (혹은 delay) $\tau=0$에서 최대값을 갖는다. 이 함수는 음성신호의 피치 (pitch 혹은 fundamental frequency), 즉 기본 주파수를 측정한다. 일반적으로 음성신호는 시간에 따라 변하는 준주기적인 신호이기에 그 주기, 즉 피치를 추정하는 것이 매우 어렵다고 알려져 있다. 본 논문에서는 피치의 측정에 ACF를 사용한다.

$$R(\tau)=\sum_{n=0}^{N-1} s(n)s(n+\tau) \tag{1}$$

Autocorrelation 함수의 계산에 곱셈이 포함되지만, 계산의 규칙적 형태 때문에 실시간으로 구현하기 쉽다. 따라서 위상 왜곡이 발생할 수 있는 신호의 피치를 탐지하는데 좋은 성능을 발휘한다.[2,3]

그림 1은 유성음의 신호를 보여준다. 이 신호는 16 kHz로 샘플링한 512 샘플을 한 프레임으로 하였다. 이에 대하여 ACF를 구한 것을 그림 2에 보였다.

ACF을 보면 그림 2와 같이 x축의 512샘플 기준 좌우 대칭의 그래프가 나타난다. 이는 주기적 성질을 갖는 그래프의 특징이다. 그림에 보인것과 같이 피치 주기는 가장 큰 피크부터 두 번째 큰 피크까지의 거리이다.[4]

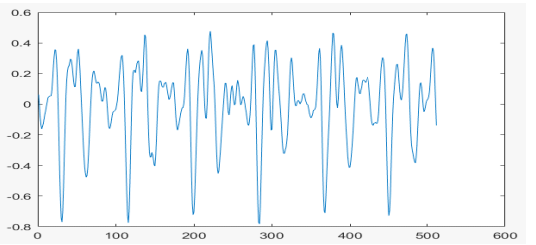

그림 1. 유성음 파형
Fig. 1. Waveform of Voiced Signal

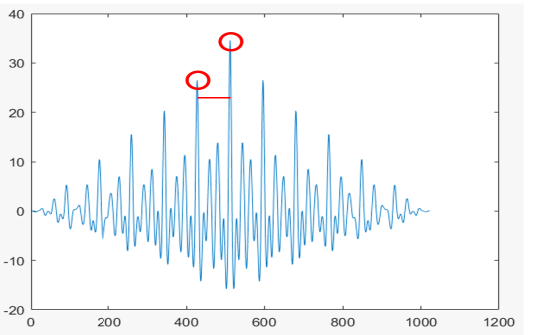

그림 2. 유성음의 ACF
Fig. 2. ACF of Voiced Signal

2.2 AMDF

ACF가 음성신호의 유사성을 찾아서 피치를 검출하는 것에 반해 AMDF는 신호의 차이로 피치를 검출한다. AMDF는 신호의 차이와 크기를 계산하므로, 그 연산량이 ACF에 비해 적고 동작 시간이 빨라서 실시간 처리에 적합하다고 할 수 있다. AMDF는 아래의 식 (2)로 정의된다.

$$D(m)=\frac{1}{N}\sum_{n=0}^{N-1-m}|s(n)-s(n+m)|\,(0\le m\le M_0) \qquad (2)$$

N샘플 프레임에서 $s(n)$은 분석하는 음성신호이고 $s(n+m)$은 m 샘플만큼 이동한 음성신호이다. 이 식에서 신호의 피치가 되는 지연 m에서 AMDF 함수가 최솟값을 갖는다. 그러므로 AMDF의 최솟값이 생기는 m의 값에서부터 $m=0$까지의 거리를 피치로 정의할 수 있다.[5,6]

그림 3은 앞의 그림 1의 유성음에 대해 AMDF를 계산한 것을 보여준다.

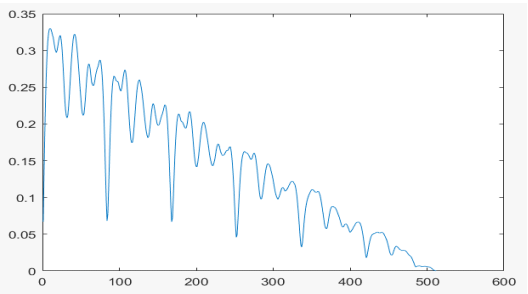

그림 3. 유성음의 AMDF
Fig. 3. AMDF of Voiced Signal

Ⅲ. Linear Predictive Coding (LPC)

3.1 LPC

임의의 시간 n에서의 음성신호는 과거의 $n-1$까지의 음성신호들로 선형조합으로 근사할 수 있다. 음성신호의 생성에서 중요한 기관이 성도 (vocal tract)인데, LPC는 이 성도를 선형화하여 음성 신호의 생성을 모델링한 것이라고 할 수 있다. 시변 디지털 필터의 출력 $s[n]$은 입력이 유성음이면, 즉 여기신호가 성대의 진동 (피치)가 되고 이는 성도를 지나면서 여러 소리를 만들어 낸다. 반면에 무성음일 경우에는 랜덤 여기신호가 성도를 따라서 소리를 생성한다. 성도의 반응을 시간에 따라 변하는 시스템(=필터)으로 보고 이의 필터 계수를 음성생성의 파라미터로 볼 수 있다. 일반적으로 이산 시스템은 차분방정식의 형태로 나타낼 수 있으므로, 이 시스템에서 계수들을 구하는 것은 성도의 특징을 알 수 있는 성도 시스템의 필터 계수를 추출 해내는 과정이 된다.[2] 정상상태 시스템 (steady state system) 함수로 이루어진 시변 디지털 필터의 전달함수는 식 (3)과 같이 pole-zero 시스템으로 나타낼 수 있다. 식 (3)을 차분방정식으로 표현하면 식 (4)와 같이 표현된다. 따라서 LPC 분석 문제 - 식 (4)의 해답을 구하는 것은 신호의 측정값이 주어지면, LPC 파라미터 - 필터의 계수 - a_j를 구하는 것이 된다.[2]

그림 4는 16kHz로 샘플링한 음성 데이터에서 유성음 프레임에 대하여 LPC 스펙트럼을 구한 것이다.

$$H(z)=\frac{S(z)}{X(z)}=\frac{G(1-\sum_{j=1}^{M}b_jz^{-1})}{1-\sum_{i=1}^{N}a_iz^{-i}} \qquad (3)$$

$$s(n)=Gu(n)+\sum_{j=1}^{p}a_js(n-j) \qquad (4)$$

그림에서 빨간색 원으로 표시한 것과 같이 4개의 peaks가 보인다. 이에 상응하는 주파수를 포먼트 (Formant) 주파수라고 한다. 이 경우에는 4개의 포먼트 주파수가 있다.[7] 모음식별에서 첫 번째 포먼트와 두 번째 포먼트가 가장 크게 관련되며, 첫 번째 포먼트는 혀의 구강 내에서 높낮이에 따른 인두강 부

피와 관련이 있고 고모음보다 저모음에서 더 높다. 두 번째 포먼트는 혀의 앞뒤 위치에 따른 구강 길이에 영향을 받고 후설모음 보다 전설 모음에서 더 높다. 세 번째 포먼트 주파수는 비음의 특성을 보인다. 일반적으로 LPC 분석에서 방정식을 10차로 한다. 그래서 이론상 포먼트 주파수는 5개까지 나타나지만, 분석하는 데이터에 따라서 보통 2~4개의 포먼트 주파수가 검출된다.[8]

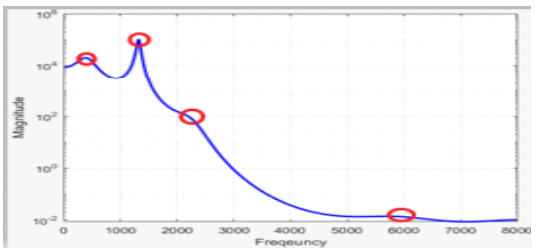

그림 4. LPC 스펙트럼
Fig. 4. LPC Spectrum

3.2 Split-LPC

앞에서 언급한 것과 같이 10차의 LPC 분석에서 2~4개의 포먼트 주파수가 검출되는데, 이는 LPC 계수를 구하는 과정이 10차의 방정식의 근들이 (poles) 서로 연관되어 영향을 주고 있는 상호간섭(interaction)이 일어나는 것으로 예측할 수 있다. 그래서 poles 사이의 상호간섭을 줄여주면 이론적인 포먼트의 수가 나타날 것으로 기대하여서 이 방법을 제시한다.[9]

[9]에서 10차의 LPC 방정식을 5개의 2차 방정식으로 분리하여서 직렬로 연결하였다. 차수가 낮은 방정식의 직렬연결은 합성필터의 poles들의 상호간섭을 줄일 수 있다. [9]에서 식(5)와 같이 구현하였다.

$$H(z)=\frac{1}{A(z)}=\frac{1}{1+\sum_{k=1}^{10}a_k z^{-k}}=\frac{1}{\prod_{k=1}^{5}(1-p_k z^{-1})(1-p_k^{*} z^{-1})} \tag{5}$$

여기서 a_k는 예측계수이고 p_k와 p_k^{*}는 $A(z)$의 근 (pole)과 켤레 복소근이다.[9]

그림 5는 전달함수 $H(z)$의 분모의 다항식 $A(z)$의 차수가 10차일 때의 스펙트럼을 보인 것이다. 이 그림에서 보면 3개의 선명한 포먼트 주파수가 있음을 알 수 있다. 그러나 3.2 kHz 대역에서의 포먼트 주파수는 선명하지 않은 것을 알 수 있다.

그림 6에서부터 그림 9까지 에서 Split-LPC 방법으로 구한 포먼트 주파수를 보였다.

이와 같이 Split-LPC 방법은 기존의 LPC 분석이 찾아내지 못하는 부분들을 선명하게 해 준다.[9]

기존의 LPC와 Split-LPC의 포먼트 주파수의 위치가 조금씩 다른 이유는 10차 방정식을 한 번에 처리한 경우, 필터를 통과해 연산이 이루어지는 과정에서

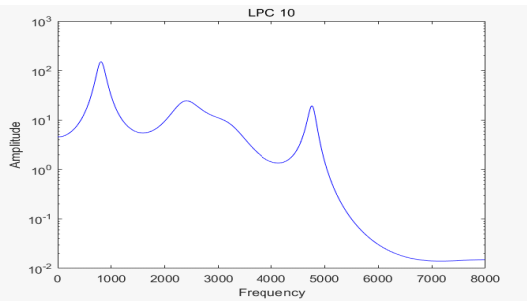

그림 5. 10차 LPC 스펙트럼
Fig. 5. LPC Spectrum of 10^{th}-order of $A(z)$

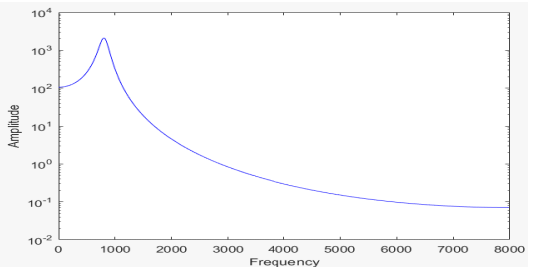

그림 6. Split-LPC의 첫 번째 포먼트 주파수
Fig. 6. 1st Formant Frequency of Split-LPC

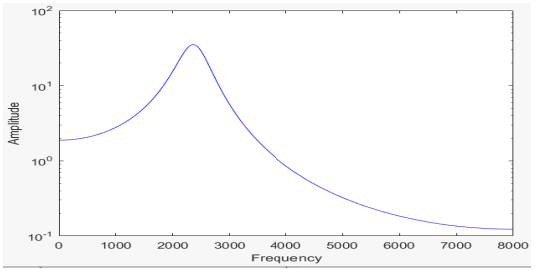

그림 7. Split-LPC의 두 번째 포먼트 주파수
Fig. 7. 2nd Formant Frequency of Split-LPC

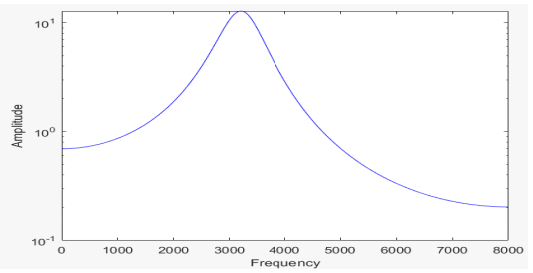

그림 8. Split-LPC의 세 번째 포먼트 주파수
Fig. 8. 3rd Formant Frequency of Split-LPC

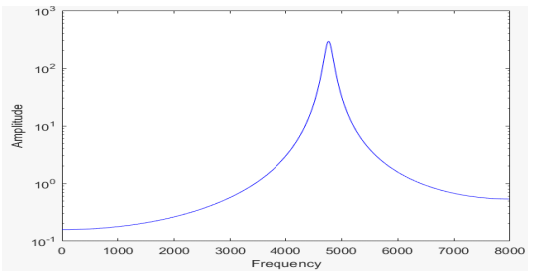

그림 9. Split-LPC의 네 번째 포먼트 주파수
Fig. 9. 4th Formant Frequency of Split-LPC

폴과 폴 사이에서 상호간섭이 일어나게 되기 때문이다. 그러므로 2차 방정식의 연산만으로 포먼트 주파수를 구하는 Split-LPC의 분석이 상호간섭을 약간은 배제할 수 있어서 그 정확도가 높다고 판단할 수있다.[9]

Ⅳ. Voiced/Unvoiced Signal Discriminating Algorithm

ZCR은 신호가 0을 지나는 횟수를 프레임으로 나누어 비교한다. 즉, 프레임당 0을 지나는 비율이다. 식 (6)으로 정의할 수 있다.

$$ZC=\sum_{t=1}^{T-1} counts_t s_{t-1} < 0 \quad (6)$$

샘플s의 t번째 값과 t-1번째 값의 곱이 0 미만이라면 두 값의 부호가 다르다는 뜻으로 zero crossing이다.[10] 유성음은 준주기적인 신호로써 무성음에 비해 0을 지나는 횟수가 적다. 반대로 무성음은 주기성이 없어 0을 지나는 횟수가 유성음에 비해 많다.[10]

STE는 시간의 변화에 따른 에너지의 값을 구하는 것으로 에너지의 식에 해밍 윈도우를 곱해줌으로써 구한다. 식 (7)으로 표현된다.

$$E=\sum_{m=-INF}^{INF} x^2(m)h(n-m) \quad (7)$$

짧은 시간동안 데이터의 제곱을 적분하는 것과 같다. ZCR과 반대로 유성음은 신호의 크기가 크며 0을 지나는 횟수도 적기 때문에 STE는 무성음에 비해 높게 나온다. 무성음은 신호의 크기가 유성음보다 대체로 작으며 0을 지나는 횟수가 많기 때문에 STE는 유성음에 비해 작게 나온다.

무성음일 때 ZCR이 높게 나타나고 STE가 낮게 나타나며 유성음일 때 ZCR이 낮게 나타나고 STE가 높게 나타난다. 본 논문은 유무성음 판별을 위해 ZCR은 입력 음성 데이터의 최대 ZCR의 60%이하, STE는 음성 데이터의 최대 STE의 20%이상인 구간을 유성음이라 판별하였다.

Ⅴ. Experiments

본 논문에서 사용한 데이터는 한국어 감정표현 언어들로 공명음을 포함한 어휘들로 구성하였고, 처음과 마지막에 채움 문장을 두어서 녹음에서 화자의 부자연스러움을 최대한 배제하였다. 이런 구성을 가진 제시문을 남자 3명과 여자 2명의 화자의 음성을 녹음하였고 이 음성 데이터를 16kHz의 주파수로 샘플링하여 본 실험을 진행하였다. 이는 제한된 환경으로 생각할 수 있다. 본 논문에서는 "우리들만 먼저 올라가자"의 청유형 문장을 사용하였고, 이 중에서 "올라가자"를 중심으로 분석하였으며 감정에서 가장 큰 차이를 보이는 기쁨과 슬픔 두 감정을 분류하였다. 실험의 진행은 아래의 단계로 수행하였다.

1. 데이터를 512샘플 단위로 프레이밍 한다.
2. 다음 프레이밍 데이터는 첫 프레이밍 데이터에서 256샘플만큼 우측으로 이동시킨 후 프레이밍 한다.
3. 총 프레임개수는 $2*\frac{\text{전체 데이터 } N}{\text{프레임 사이즈 } 512}-1$ 이다.
4. 각각의 프레임 당 항목별로 평균값을 내어 파라미터를 추출한다.

그림 10은 프레임당 피치 주기와 ZCR의 평균값을 그래프로 나타낸 것이다. 이 그래프에서 ZCR이 일정 수준 이상 높아지면 피치 주기가 하락하는 것을 볼 수 있다. 모든 데이터는 16kHz로 샘플링 되었기 때문에 샘플간의 거리는 0.0625 (ms)이다. 따라서 기본 피치 주기(샘플거리)에 0.0625 (ms)를 곱하여 피치 주기를 시간으로 나타내었다. 남자와 여자의 피치주기는 확연히 차이가 나며 남자가 더 긴 것을 확인할 수 있다. 또한, 대체로 슬픔이 기쁨에 비해 피치 주기가 긴 것을 확인할 수 있다.

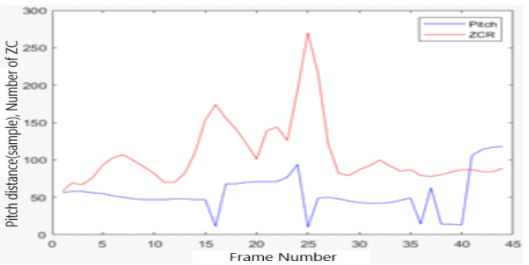

그림 10. ZCR과 피치주기의 관계
Fig. 10. Relationship between pitch periods and ZCR

Ⅵ. SVM 분류기의 제안

SVM은 두 카테고리 중 어느 하나에 속한 데이터 집합이 주어졌을 때, 주어진 데이터 집합을 바탕으로 새로운 데이터가 어느 카테고리에 속하는지 판단하는 이진 선형분류 모델이다.

위 그림 11처럼 두 부류 사이의 여백이 가장 넓어지면 잘 분류하였다고 할 수 있다. 즉 margin이 최대화 되면 좋은 분류이다. 두 카테고리의 data set의 최외각에 있는 샘플들을 Support Vector라고 하며 이 벡

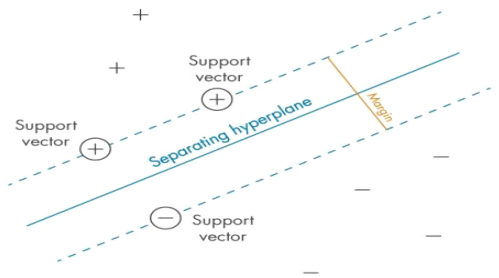

그림 11. SVM의 마진의 정의
Fig. 11. Definition of margin for SVM

터들은 margin을 구하는 데 사용되기 때문에 중요한 벡터이다. 즉, 성능이 우수한 SVM을 만들기 위해서는 Support Vector를 통해 margin이 최대가 되는 직선을 찾아야 하고 이를 결정 초평면(Decision Hyperplane)이라고 한다.[11,12]

결정 초평면을 구하기 위해선 초평면에 수직인 가중치벡터 w와 상수 b를 구해야 한다.

$$d(x) = \vec{w} \cdot \vec{x} + b = 0 \tag{8}$$

margin의 길이 width= $\frac{2}{\| \vec{w} \|}$ 이므로 width를 최대화하는 조건부 최적화 문제로 해석된다. 해석에는 라그랑주 승수법이 사용된다. 따라서 식(9)의 라그랑주 함수로 표현할 수 있다. α_i는 라그랑주 승수이며 y_i는 훈련 집합이고 N은 훈련샘플의 개수이다.

$$\ell(w,b,\alpha) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{N} \alpha_i [y_i (w \cdot x_i + b) - 1] \tag{9}$$

라그랑주 승수법은 최적점이 되기 위한 조건을 찾는 방법, 즉 최적해의 필요조건을 찾는 방법이다.

$$\ell(w,b,\alpha) = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j = \ell(\alpha) \tag{10}$$

최종적으로 식(10)을 통해 라그랑제 승수 α값을 구할 수 있으며 그로 인해 w,b를 알 수 있다.[13,14]

앞의 섹션Ⅴ에서 보인 표 1과 2의 파라미터들을 SVM 분류기를 이용하여서 감정을 분류하였다. 앞서 설명했듯이 남자와 여자는 피치 주기에서 커다란 차이를 보였기에 이 파라미터를 사용하여 남자와 여자를 분류하였다.

그림 12와 13에서 피치 주기와 첫 번째 포먼트 주파수와 두 번째 포먼트 주파수를 이용하여 성별 분류를 하였다.

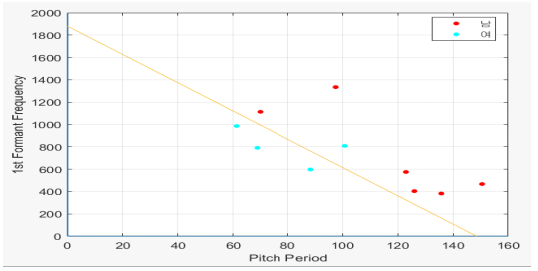

그림 12. 성별 분류 (피치 주기-첫 번째 포먼트 주파수)
Fig. 12. Gender Classification (Pitch period-1st Formant Frequency)

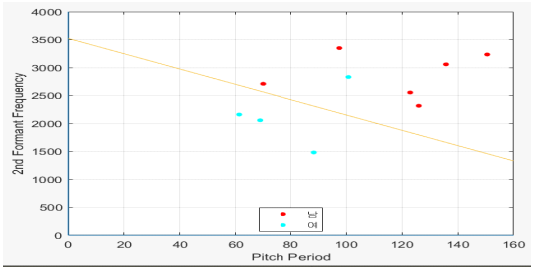

그림 13. 성별 분류 (피치 주기-두 번째 포먼트 주파수)
Fig. 13. Gender Classification (Pitch period-2nd Formant Frequency)

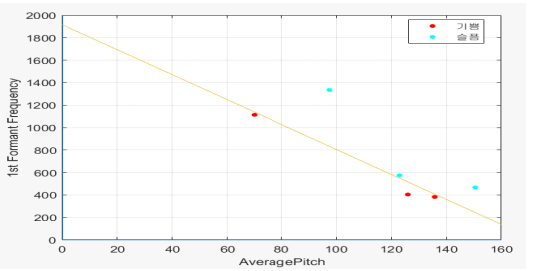

그림 14. 남성의 감정 분류 (피치 주기-첫 번째 포먼트 주파수)
Fig. 14. Emotion Classification of Male (Pitch period-1st Formant Frequency)

남자인 경우, 기쁨과 슬픔의 감정의 분류는 "피치 주기-첫 번째 포먼트 주파수"와 "피치 주기-두 번째/첫 번째 포먼트 주파수"의 두 가지 경우가 높은 정확도의 결과를 나타내었다. "피치 주기- 두 번째 포먼트 주파수"의 경우 정확도가 낮았지만, 두 번째 포먼트 주파수간 차이가 첫 번째 포먼트 주파수간 차이보다 작기 때문에 두 포먼트 주파수 사이의 상대비를 그림 15에서 보였다. 그 결과 오차 없는 이상적인 결과를 나타내었다.

여자의 경우는 기쁨과 슬픔의 감정 분류를 "피치

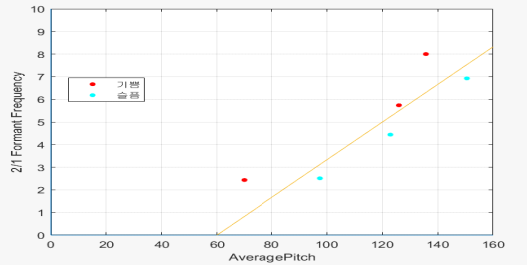

그림 15. 남성의 감정 분류 (피치 주기-두 번째/첫 번째 포먼트 주파수)
Fig. 15. Emotion Classification of Male (Pitch period - 2nd/1st Formant Frequency)

주기-첫 번째 포먼트 주파수"와 "피치 주기-두 번째 포먼트 주파수"로 높은 정확도의 결과를 얻었다. 이 결과는 그림 16과 17에서 보여준다.

그림 12에서 그림 17까지의 SVM 분류기를 이용하여 감정분류의 과정을 보였다. 본 논문에 보인 결과들은 다른 파라미터들의 실험 결과보다 우수함을 보인 것들이다. 그래서 아래와 같이 SVM 분류기를 활용한 감정분류 알고리즘을 제안한다.

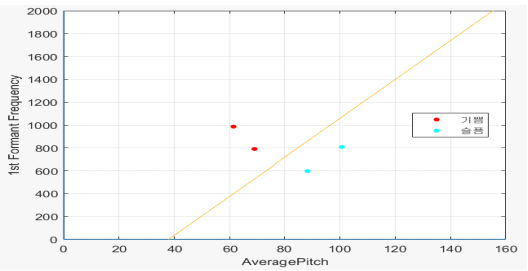

그림 16. 여성의 감정 분류 (피치 주기-첫 번째 포먼트 주파수)
Fig. 16. Emotion Classification of Female (Pitch period-1st Formant Frequency)

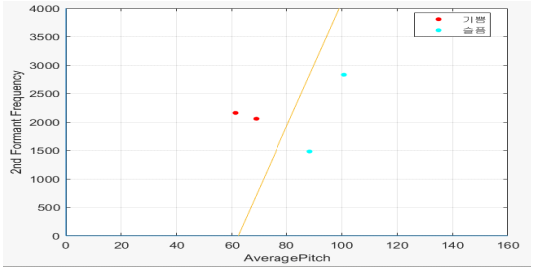

그림 17. 여성의 감정 분류 (피치 주기-두 번째 포먼트 주파수)
Fig. 17.Emotion Classification of Female (Pitch period-2nd Formant Frequency)

6.1 성별 분류 시행

피치 주기와 첫 번째 또는 두 번째 포먼트 주파수의 파라미터들을 이용하여 남자 혹은 여자로 분류한다.

6.2 남성인 경우

피치 주기와 첫 번째 포먼트 주파수로 감정 분류를 시행한 후에 피치 주기와 첫 번째 와 두 번째 포먼트 주파수의 상대비 (F1/F2)의 파라미터들로 최종적 감정 분류를 한다.

6.3 여성의 경우

피치 주기와 첫 번째 포먼트 또는 두 번째 포먼트 주파수를 파라미터로 하여서 감정을 분류한다.

Ⅶ. Conclusion

감정, 성별에 따른 음성 데이터의 피치 주기 및 포먼트 주파수 차이를 확인하였다. 본 논문에서 개발한 유 무성음 판별 알고리즘의 정확도가 적절히 높았다. 성별을 분류하여 감정분류의 정확도를 개선하는 방식을 제안하였다. 남성의 피치 주기는 약 5~10ms, 여성의 피치 주기는 약 3.5~6.5ms로 남성의 피치 주기가 더 길다는 것을 확인하였다. 그리고 슬픔 감정은 남성 6~9.5ms, 여성 5.5~6.5ms 기쁨 감정은 남성 4~8ms, 여성 3.5~4.5ms로 확인되었다. 여기에 포먼트 주파수를 파라미터로 활용하여 정확한 판단을 할 수 있었다. 포먼트 주파수의 측정은 pole의 상호간섭을 감소시켜 정확도를 높이는 Split-LPC 방식을 제안하였다.

SVM 분류의 정확도를 높이기 위해 추출된 파라미터를 이용하여 SVM으로 분류를 시행할 때 margin의 크기보다 에러가 적게 나오는 것에 가중치를 두었다. 가중치가 커질수록 에러율이 작아지며 margin이 작아지고, 가중치가 작아질수록 에러율이 커지며 margin의 크기가 커진다. 즉, 현재의 데이터만을 기준으로 분류를 시행한다면 가중치를 크게 설정하여 에러율을 줄이는 것이 성능에 좋다. 반대로 추후에 많은 양의 데이터 입력을 예상한다면 가중치를 작게 설정하여 margin을 크게 설정하는 것이 좋다.

본 논문에 보인 시뮬레이션 결과 이외에도 피치 주기 - 세 번째 포먼트 주파수, 첫 번째 포먼트 주파수 - 두 번째 주파수 등 여러 경우의 시뮬레이션을 진행하였지만 에러율이 33%~66%의 높은 수치로 발생하였기 때문에 가장 정확도가 높은 실험들을 개재하였다. 제한된 환경에서 감정 분류를 시행할 수 있으며,

이로인해 네트워크 연결이 어려운 지역, 낮은 성능의 기기 또는 데이터가 부족한 환경에서도 사용할 수 있다. 제안된 알고리즘으로 더 높은 정확도의 파라미터 추출과 분류 방식으로 성능이 향상되었다.

본 논문에서 제안한 알고리즘은 간단한 분류기인 SVM을 이용하였기에 향후 CNN, RNN과 같은 뉴럴 네트워크를 활용한다면 성능의 향상을 기대할 수 있다. 그리고 기쁨 및 슬픔 두 가지 감정뿐 아니라 평상시, 화남, 놀람 등의 여러 가지 감정을 분류할 수 있을 것이다. 본 연구를 AI 비서 및 AI 스피커에 적용하여 사용자의 감정에 따른 서비스를 제공한다면 현재까지 상용화되지 않던 사용자 최적화 방향으로 큰 발전이 기대된다.

References

[1] G. E. Jo and S. I. Kim, "A study on user experience of artificial intelligence speaker," *J. Korea Convergence Soc.*, vol. 9, no. 8, pp. 127-133, 2018.

[2] A. M. Kondoz, *Digital Speech(Coding for Low Bit Rate Communications Systems)*, 1st Ed., WILEY, 1995.

[3] S. A. So, K. H. Lee, K. B. You, H. Y. Lim, and J. S. Park, "A study of peak finding algorithms for the autocorrelation function of speech signal," *J. The Korea Soc. Comput. and Info.*, vol. 21, no. 12, pp. 131-137, Dec. 2016.

[4] S. S. Upadhya, "Pitch detection in time and frequency domain," in *ICCICT*, Mumbai, India, Oct. 2012.

[5] B. Kotnik, H. Hoge, and Z. Kacic, "Evaluation of pitch detection algorithms in adverse conditions," in *Proc. 3rd Int. Conf. Speech Prosody*, pp. 149-152, Dresden, Germany, 2006.

[6] S. So, K. H. Lee, K. B. You, H. Y. Lim, and J. Park, "A study of the pitch estimation algorithms of speech signal by using average magnitude difference function," *Asia-pacific J. Multimedia Services Convergent with Art, Humanities, and Sociology*, vol. 7, no. 4, pp. 235-242, Apr. 2017.

[7] S. D. Heo and H. R. Kang, "Formant frequency changes of female voice /a/, /i/, /u/ in real ear," *Phonetic and Speech Sci.*, vol. 9, no. 1, pp. 49-53, 2017.

[8] I. V. McLoughlin, "A review of line spectral pairs," *Preprint submitted to Elsevier*, Sep. 2007.

[9] K. B. You and K. H. Lee "A study on splitting LPC synthesis filter," *Info. Technol. Convergence*, LNEE, vol. 253, pp. 1003-1009, Springer, Jul. 2013.

[10] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," *ASEE*, pp. 1-7, 2008.

[11] E. Garcia-Gonzalo, Z. Fernandez-Muniz, and P. J. Garcia Nieto, "Hard-Rock stability analysis for span design in entry-type excavations with learning classifiers," *Materials*, vol. 9, no. 7, 2016.

[12] K. S. Durgesh and B. Lekha, "Data classfication using support vector machine," *J. Theoretical and Applied Info. Technol.*," vol. 12, no. 1, pp. 1-7, 2010.

[13] E. Osuna, R. Freund, and F. Girosi, "Support vector machines: Training and applications," *CSAIL*, A.I. no. 1602, MIT, 1997.

염 정 석 (Jeong-seok Yeom)

2014년 2월~현재 : 숭실대학교 전자정보공학부 (IT융합전공) 재학중

<관심분야> 인공지능, 신호처리, 전자공학, 통신공학

[ORCID: 0000-0003-2199-2320]

유 광 복 (Kwang-Bock You)

1998년 5월 : Stevens Inst. of Tech. 공학박사
2010년 9월~현재 : 숭실대학교 전자정보공학부 교수
<관심분야> 음성신호처리, 샘플링 정리, 무선 통신, 인공지능

[ORCID: 0000-0002-3311-2418]

장 경 남 (Kyungnam Jang)

1998년 2월 : 숭실대학교 국어국문학과 문학박사
2001년 3월~현재 : 숭실대학교 국어국문학과 교수
<관심분야> 한국 고전소설, 고전과 문화콘텐츠

[ORCID: 0000-0001-7028-3876]