효과적인 감정인식을 위한 음성 특징 벡터 생성

심인우*, 한의환*, 차형태*
* 숭실대학교 전자공학과

e-mail : tladlsdn@naver.com

Generating Speech feature vectors for Effective Emotional Recognition

In-woo Sim*, Eui Hwan Han*, Hyung Tai Cha*
Dept of Electric Engineering, Soongsil University

요 약

본 논문에서는 효과적인 감정인식을 위한 효과적인 특징 벡터를 생성한다. 이를 위해서 음성 데이터셋 RAVDESS를 이용하였으며, 그 중 neutral, calm, happy, sad 총 4가지 감정을 나타내는 음성 신호를 사용하였다. 본 논문에서는 기존에 감정인식에 사용되는 MFCC1~13차 계수와 pitch, ZCR, peakenergy 중에서 효과적인 특징을 추출하기 위해 클래스 간, 클래스 내 분산의 비를 이용하였다. 실험결과 감정인식에 사용되는 특징 벡터들 중 peakenergy, pitch, MFCC2, MFCC3, MFCC4, MFCC12, MFCC13이 효과적임을 확인하였다.

1. 서론

최근 많은 기업들에서 가정용 음성 비서, 휴대폰 음성 비서 등에 관심을 보이며 개발에 힘쓰고 있다[1]. 음성 비서는 화자의 음성을 인식하고 음성에 담긴 정보를 토대로 화자에게 정보를 제공하거나 대답을 해주는 AI이다. 음성에는 나이, 성별, 감정 등과 같은 정보가 담겨있다. 그 중 감정은 화자의 심리상태를 나타내며, 이 정보는 AI가 화자에게 답변을 해줄 때 중요한 포인트가 된다[1]. 최근들어 음성 특징에서 감정분석을 위한 효과적인 특징을 생성하는 연구가 활발히 진행되고 있다[2]. 음성에 담겨있는 감정을 구분하기 위하여 기존 연구에서는 Mel Frequency Cepstral Coefficients(MFCC), pitch, Zero Crossing Rate(ZCR), Integral Absolute Value(IAV), Linear Predictive Coding(LPC), Tempo와 같은 다양한 특징 벡터를 사용했다[3].

따라서, 본 논문에서는 감정인식을 위해 클래스 간, 클래스 내 분산의 비율을 이용하여 음성 특징 중 감정인식에 효과적인 특징을 생성한다.

2. 기존의 연구

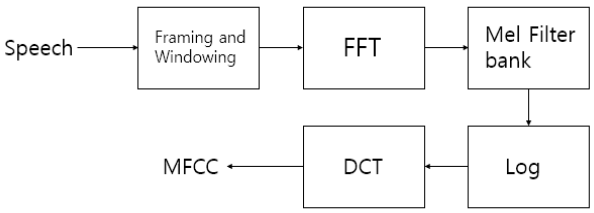
본 장에서는 본 논문에서 사용될 음성 특징 벡터와 기존의 연구에 대해 소개한다.

2.1 음성 특징 벡터

기존 연구에서는 음성감정분석을 위하여 다양한 특징 벡터들을 사용하였다. 대표적으로 사용되는 음성 특징 벡터에는 MFCC, pitch, ZCR, peakenergy등이 있다[2]. 본 논문에서는 이러한 특징 중 감정인식에 효과적인 특징을 찾는다. 각 특징에 대한 설명은 아래와 같다.

2.1.1 MFCC

Mel Frequency Cepstral Coefficients의 약어로 인간의 귀가 낮은 주파수 영역에서는 민감하고 높은 주파수 영역에서는 둔감한 비선형적인 주파수 특성을 나타내는 단위인 Mel 스케일을 기반으로 한 음성 특징이다[2]. 본 논문에서는 MFCC1~13차 계수를 추출하여 특징 벡터 비교에 이용하였다. MFCC는 (그림1)과 같이 구할 수 있다.



(그림 1) MFCC 추출 과정

입력되는 음성 신호를 작은 크기의 프레임으로 자른 후 각 프레임에 대하여 power spectrum을 계산한다. 구해진 power spectrum에 Mel Filter bank를 적용하고, 구해진 값들에 Log를 취한다. 마지막으로 DCT를 취하면 MFCC가 구해진다.

MFCC는 실제로 23개의 값이 추출되지만, 음성인식에 크게 기여하지 못하는 계수를 제외하면 13개가 남는다[4].

때문에 본 연구에서는 13차까지의 MFCC만 사용한다.

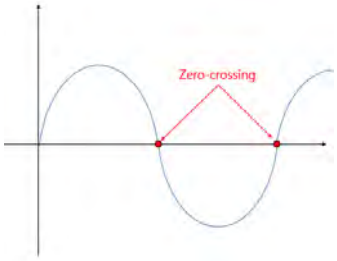
2.1.2 Pitch

Pitch의 검출은 소리의 높낮이를 나타내는 고유 주파수를 찾아내는 것이다. pitch검출에는 캡스트럼 방법과 YIN 기반 pitch추출방법이 존재한다. YIN기반 pitch추출방법은 기존의 캡스트럼 방법에 비해 계산량이 적고, 고조파 성분이 검출될 가능성이 낮다[5]. 따라서 본 논문에서는 성능이 더 좋은 YIN기반 pitch추출 방법을 사용하여 pitch를 검출한다.

2.1.3 Zero-Crossing Rate(ZCR)

ZCR은 신호의 파형이 0축을 지나는 횟수, 즉 신호의 부호가 바뀌는 비율이다. ZCR은 식(1)과 같이 계산된다.

$$zcr = \frac{1}{T-1}\sum_{l=1}^{T-1} 1_{R<0}(S_t S_{t-1}) \quad (1)$$



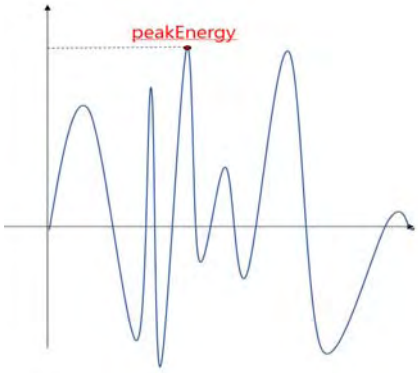
(그림 2) Zero-Crossing

2.1.4 peakenergy

peakenergy는 신호가 가지는 최댓값을 의미한다. 본 연구에서 사용되는 peakenergy의 값은 음성에서데이터 셀에 저장된 값들 중 가장 큰 값으로 정하였다.

$$peakEnergy = MAX(S(t)) \quad (2)$$

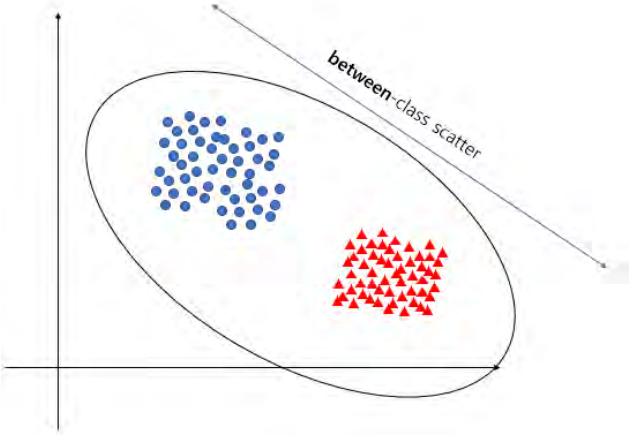
$S(t)$: 음성신호



(그림 3) peakenergy

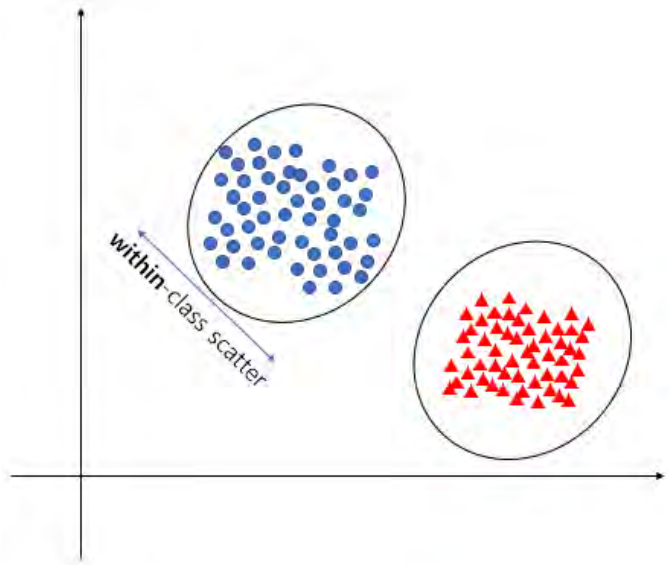
2.2 선형판별분석

선형 판별 분석은 클래스로 분류되어있는 다차원 공간의 데이터에 대해 클래스간의 분산이 최대가 되도록 하는 이론이다. 클래스 간의 분별 정보를 최대한 유지, 즉 클래스 간 분산이 최대가 되며 클래스 내 분산이 최소가 되도록 차원을 축소하여 분별이 쉬워지도록 하는 것이다[6].



(그림 4) 클래스 간 분산

클래스 간 분산 값은 표본 집단 간의 거리를 나타내며, 이 값이 클수록 두 표본 집단이 멀리 떨어져있다. 표본 집단 간의 거리가 멀수록 특징 분리는 쉬워진다.



(그림 5) 클래스 내 분산

클래스 내 분산 값은 각각의 표본 집단의 원소 간의 값들이 서로 얼마나 거리가 떨어져 있는지를 나타내며, 이 값이 작을수록 원소들이 비슷한 값을 가진다. 표본 집단의 원소들 간의 값이 비슷할수록 특징 분리는 쉬워진다.

3.제안하는 알고리즘

본 장에서는 2장에서 설명한 것 같이 클래스 내 분산과 클래스 간 분산의 비를 이용하여 효과적인 음성 특징 벡터를 생성한다.

3.1 데이터셋

본 연구에서 사용한 데이터셋은 RAVDESS Dataset[7]이고, 이에 대한 정보는 <표 1>과 같다.

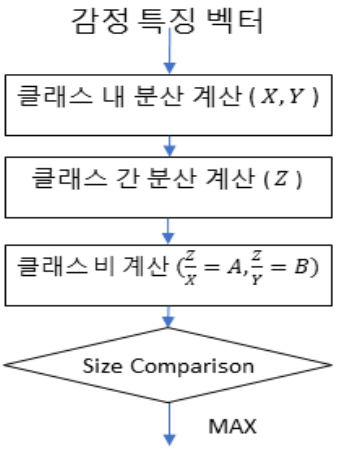
<표 1> 사용한 RAVDESS Dataset 정보

Index	Information
연기자 인원 수	15명
언어	영어(북미)
연기자 성 비	남성 : 9명, 여성 : 6명
감정 카테고리	neutral, calm, happy, sad
사용된 문장	"Kids are talking by the door"

감정 카테고리로는 중립(neutral), 평온함(calm), 즐거움(happy), 슬픔(sad) 4가지로 정의한다. 특징벡터 추출과 분석을 위해 쓰인 데이터는 각 감정 당 15개의 데이터로 총 60개이다. 남성 음성과 여성 음성의 비율은 9 대 6의 비율로 사용하였다.

3.2 제안하는 알고리즘

본 논문에서는 다음과 같은 알고리즘을 통하여 음성 특징 벡터를 비교하고, 효과적인 음성 특징 벡터를 생성한다.



(그림 6) 분리도 계산 흐름도

첫 번째, 두 개의 클래스에 대해서 클래스 간의 분산과 각각의 클래스 내 분산을 구한다.

클래스 내 분산은 표본 집단들의 밀집도를 나타내는데 이 값이 작을수록 밀집도가 크다. 클래스 간 분산은 두 클래스 간 밀집도로 값이 클수록 두 클래스간의 분리도가 크다.

두 번째, 클래스 간의 분산을 클래스 내 분산으로 나누어 각 감정에 대한 클래스 비를 구한다. 클래스 간 분산을 클래스 내 분산으로 나눈 값, 즉 클래스 간, 클래스 내 분산의 비가 클수록 두 특징 간의 분리가 쉬워진다.

마지막으로 두 개의 클래스 비 중 가장 큰 값을 선택하여 분리도로 결정한다.

그 중 두 감정을 분리할 때 적합한 음성 특징 벡터는 구해진 분리도 중 높은 값을 가진 특징 벡터를 선택한다.

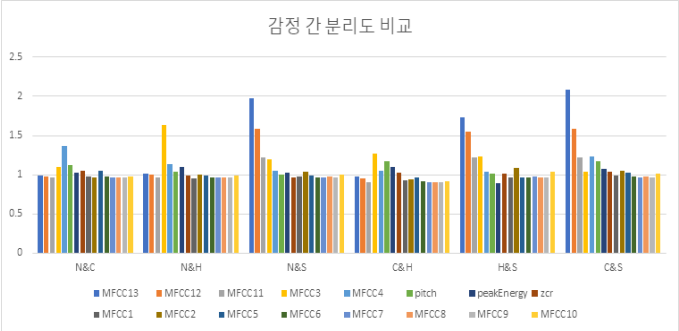
4. 실험 결과

<표 2> 감정 간 특징 벡터에 따른 분리도

	N&C	N&H	N&S	C&H	C&S	H&S
peak Energy	0.91878	1.10095	1.02439	1.0948	1.0699	0.8958
	1.02903	0.87587	0.95748	0.77765	0.89287	1.05246
zcr	0.8836	0.98978	0.96339	1.02099	1.03458	0.92712
	1.04441	0.90885	0.96369	0.79314	0.87554	1.00999
pitch	1.12226	0.89902	0.93774	0.88047	0.9656	1.01067
	0.98345	1.04341	0.99533	1.16611	1.16955	0.92428
MFCC1	0.98205	0.95865	0.95817	0.90246	0.96305	0.96498
	0.97483	0.97318	0.97379	0.92294	0.98601	0.96606
MFCC2	0.96588	1.00643	0.91048	0.94411	0.91575	0.87766
	0.96675	0.92783	1.04033	0.8696	1.0454	1.08778
MFCC3	0.87444	0.76775	0.83738	0.74912	0.90951	1.22993
	1.09604	1.63404	1.19516	1.27202	1.03565	0.82479
MFCC4	1.36784	1.13369	1.04748	0.81942	0.83923	0.90649
	0.7936	0.8460	0.89553	1.05393	1.23664	1.03853
MFCC5	1.05075	0.98323	0.98801	0.85596	0.91742	0.96942
	0.90738	0.95317	0.95022	0.96091	1.02175	0.96175
MFCC6	0.95631	0.96504	0.96272	0.91258	0.97243	0.96306
	0.97618	0.96607	0.96896	0.89497	0.95881	0.96826
MFCC7	0.96302	0.96631	0.96337	0.90641	0.96552	0.96265
	0.96809	0.96472	0.9701	0.90018	0.96717	0.97097
MFCC8	0.96579	0.96407	0.96326	0.90164	0.96303	0.96432
	0.96525	0.96701	0.97176	0.90489	0.97207	0.96987
MFCC9	0.96713	0.96988	0.9709	0.90579	0.96935	0.96727
	0.96391	0.96134	0.96255	0.90081	0.96422	0.96747
MFCC10	0.97132	0.98387	0.94497	0.91466	0.94036	0.93821
	0.95982	0.95076	1.00717	0.89446	1.01426	1.03479
MFCC11	0.9672	0.96715	1.21907	0.90319	1.21886	1.21895
	0.96386	0.96391	0.93419	0.90329	0.93727	0.93723
MFCC12	0.9739	0.94548	1.58208	0.8821	1.59002	1.55241
	0.95819	1.00015	1.02136	0.9484	1.04331	0.94741
MFCC13	0.94985	0.99996	1.97059	0.97826	2.08032	1.73145
	0.9844	1.01389	1.00178	0.95707	1.02044	0.86812

감정들 간의 비교분석을 하였을 때, 감정인식에 사용되는 특징 벡터들 중 peakenergy, pitch, MFCC2, MFCC3, MFCC4, MFCC12, MFCC13 가 분리도가 높게 측정되었다.

특히, neutral&calm 에서는 MFCC4, neutral&happy 에서는 MFCC3, neutral&sad 에서는 MFCC13, calm&happy에서는 MFCC3, calm&sad에서는 MFCC13, happy&sad에서는 MFCC13이 가장 분리도가 높게 나타났다.



(그림 7) 감정 간 분리도 차트

<표 3> 감정 간의 분리도가 높은 특징 벡터

	특징 벡터
neutral&calm	MFCC4,pitch,MFCC3
neutral&happy	MFCC3,MFCC4,peakenergy
neutral&sad	MFCC13,MFCC12,MFCC11,MFCC3
calm&happy	MFCC3,pitch,peakenergy
calm&sad	MFCC13,MFCC12,MFCC4,MFCC11
happy&sad	MFCC13,MFCC12,MFCC3

<표 3>은 각 감정들 간의 비교 분석 시에 분리도가 높게 나타난 특징 벡터들이다.

sad와 다른 감정들 간의 분리도가 높게 나타난 특징 벡터는 MFCC13과 MFCC12, MFCC3이다. 이를 통해 sad와 다른 감정들을 분리할 때 이 특징들이 효과적임을 알 수가 있다. 마찬가지로, neutral의 경우에는 MFCC3이 공통적으로 분리도가 높았고, happy의 경우에도 MFCCC3이 분리도가 높았다. calm은 분리도가 높은 특징 벡터가 공통적으로 pitch가 있지만, calm&sad 에서는 pitch의 분리도가 높지 않았다.

전체적인 감정인식에서는 대부분에서 분리도가 높은 MFCC3과 MFCC4를 이용하여 효과적인 감정인식을 할 수 있다.

5.결 론

본 논문에서는 감정인식을 위한 기존에 음성 특징 벡터(MFCC, pitch, ZCR, peakenergy)중 효과적인 특징을 생성한다. 데이터셋으로는 RAVDESS Dataset의 음성 파일을 사용하였고, neutral, calm, happy, sad 총 4가지 감정으로 분류한다. 음성 벡터 특징을 생성하기 위하여 클래스 내, 클래스 간 분산의 비를 이용하였다. 분석 결과 peakenergy, pitch, MFCC2, MFCC3, MFCC4, MFCC12, MFCC13 가 분리도가 높게 측정되었고, 각 감정 별로 공통적으로 분리도가 높게 나온 특징 벡터는 sad의 경우 MFCC12, MFCC13, MFCC3, neutral과 happy의 경우 MFCC3 이었고, 각 감정을 인식할 때 더욱 좋은 효율을 내는 특징 벡터임을 확인하였다.

참고문헌

[1] 김동주, 이가람, 김호원, “AI를 활용한 음성인식 기법 및 개인비서 서비스 기술 분”, 한국통신학회 학술대회논문집, pp 213-214, 2017

[2] E. H. Han and H. T. CHA, “Performance Improvement in a Classification method by using Statistical Pre-processing”, ournal of Institute of Control, Robotics and Systems, 25(1), 2019

[3] E. H. Han and H. T. CHA, “Audio Feature Extraction for Effective Emotion Classification”, Transactions on Smart Processing and Computing, (In press), 2019

[4] 조지혁, 유호영, 차소영, 박인철, “MFCC 특징 추출을 위한 Floating-Point 비트 너비 최적화 방안”, 대한전자공학회 학술대회, pp 1194-1197, 2013

[5] 석수영, 정현열, “YIN 피치 정보를 이용한 음악 정보 검색 시스템 구현”, 멀티미디어학회논문지, 10(11), pp 1398-1406, 2017

[6] 강지수, 김희석, 홍석희, “선형판별분석을 이용한 전력 분석 기법의 성능 향상”, 정보보호학회논문지, 24(6), pp 1055-1063, 2014

[7] Livingstone SR, Russo FA, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e019639”,. 2018