음성 분석 기반 언어 인지 모델

Audio based Language Identification Model

https://github.com/joowoniese/LanguageIdentification

Ⅰ. 서 론

최근 세계화의 진전에 따라 국제적인 의사소통이 필수적인 요소로 자리 잡고 있다. 이에 따라 네이버의 파파고(Papago)와 구글 번역기와 같은 플랫폼은 사용자의 발화를 음성으로 받아 실시간으로 원하는 언어로 번역해 주는 기능으로 큰 인기를 끌고 있으며, 이 분야의 발전을 위해 많은 노력이 기울여지고 있다.

이러한 플랫폼은 여행 시나 타국인 간의 정보 교류 시 사용자들에게 큰 편리함을 제공하고 있다. 본 연구에서는 발화자의 음성을 통해 언어를 자동으로 인식하는 시스템을 제안한다. 본 연구를 통해 사용자는 단순한 실시간 번역을 넘어 발화자의 언어를 별도로 지정하지 않아도 음성을 통해 자동으로 언어를 인식할 수 있는 기능을 이용할 수 있으며, 이는 더욱 편리한 사용자 경험을 제공할 것이다.

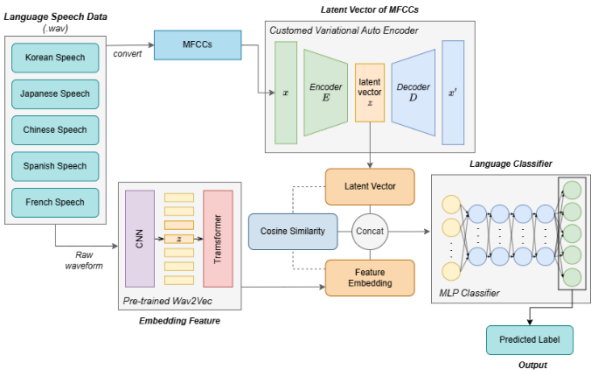


그림1. 언어 인식 모델의 전체적인 구조

본 연구에서는 음성 데이터로부터 억양, 발화 속도 등의 특징을 추출하여 발화자의 언어를 인식하는 모델을 제안한다. 그림 1에서 보이듯, 음성 데이터는 먼저 음성을 잠재 벡터로 변환하는 두 개의 모델을 거치며, 이후 이 잠재 벡터들을 입력으로 사용하는 Classifier를 통해 최종적으로 언어 분류를 수행한다.

Ⅱ. 언어 인식 모델

음성 오디오를 기반으로 언어를 분류하기 위해, 음원에서 주요 오디오 특징을 추출하고 이를 벡터화하는 과정이 필요하다. 본 연구에서는 두 가지 방식으로 벡터화를 수행하였다. 첫 번째로, 음성을 사전 학습된 wav2vec 모델에 입력하여 벡터화를 수행하였고, 두 번째로는 음원에서 직접 추출한 특징값을 개발한 Variational Autoencoder(VAE)의 Encoder를 통해 잠재 벡터를 생성하였다.

이렇게 추출된 두 개의 벡터를 결합 및 유사도를 측정하여 최종 Classifier 모델의 입력으로 사용하였으며, Classifier는 Multi-Layer Perceptron(MLP) 구조로 구축되었다. 최종적으로, Classifier는 결합된 벡터를 바탕으로 발화 언어를 분류하는 역할을 수행한다.

1.1 Wav2Vec 2.0

Wav2Vec은 음성 데이터를 효율적으로 학습하고 처리하기 위해 음성 신호의 embedding을 학습하여, 음성 인식, 언어 이해 등의 작업에서 사용할 수 있게 설계된 모델이다.

Wav2Vec은 음성 데이터를 효율적으로 학습하고 처리하기 위해 설계된 모델로, 음성 신호의 임베딩(embedding)을 학습하여 음성 인식 및 언어 이해와 같은 작업에서 활용할 수 있다. 이 모델은 Feature Encoder, Transformer Network, Quantization Module로 구성되어 있으며, 연속적인 음성 데이터를 고정 길이 벡터인 임베딩으로 변환하는 것을 목표로 한다[1].

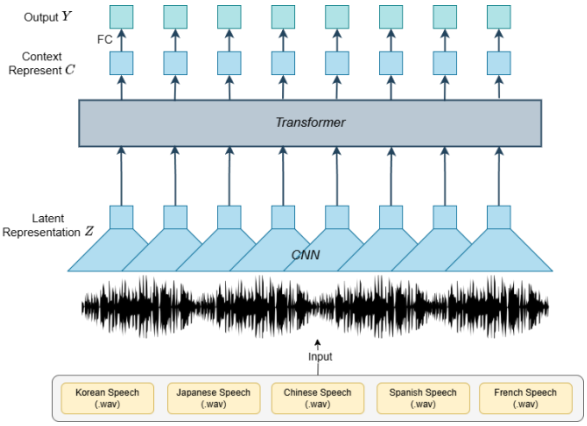


그림2. Wav2Vec 2.0 모델 구조

Wav2Vec은 원시 오디오 신호(raw waveform)를 직접 입력으로 받아, 레이블이 부족한 환경에서도 음성 데이터의 의미 있는 특징을 효과적으로

학습할 수 있다. 모델 구조는 그림 2에 제시되어 있으며, 입력된 원시 오디오 신호를 프레임 단위로 처리하고, 일부 프레임을 masking하여 학습하는 Transformer 기반의 구조이다. 이를 통해 Context 표현을 학습하여 음성 데이터의 중요한 특징을 추출한다.

CNN 기반 Encoder는 입력 데이터를 저차원 특징으로 변환하며, Self-Attention 메커니즘을 통해 프레임 간의 관계를 학습한다. Contrastive Loss는 음성 신호에서 중요한 정보를 보존하면서 노이즈와 불확실성을 제거하는 데 기여한다. 그림 4는 언어 레이블별로 벡터화된 결과를 나타내며, 각 언어 레이블에 따라 데이터 임베딩이 효과적으로 추출되었음을 보여준다.

1.2 Variational Auto Encoder

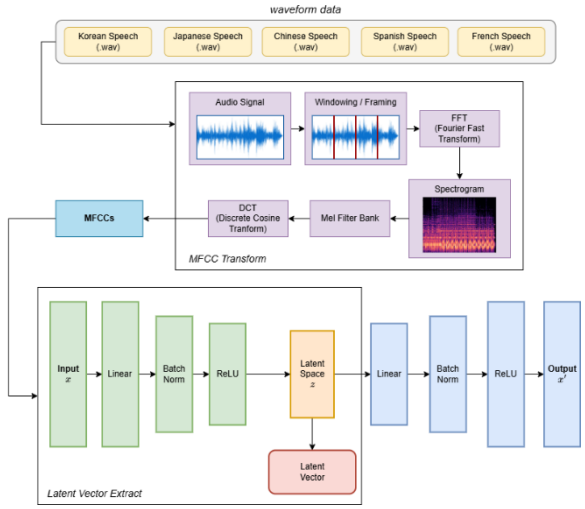


그림3. VAE 모델 구조

더욱 정교한 언어 분류를 위해, Wav2Vec으로 추출된 특징에 더하여 각 언어별 음성에서 추가적인 특징을 추출하는 과정을 도입하였다. 이를 위해 직접 개발한 Variational Autoencoder(VAE) 모델을 학습시켜 추가적인 벡터화를 수행하였다. VAE는 모델 전체를 학습한 후, 잠재 벡터를 추출하기 위해 Encoder만을 활용한다. 이렇게 추출된 잠재 벡터는 최종 Classifier의 입력 벡터 중 하나로 사용되어 분류 성능을 향상시키는 데 기여한다.

본 연구에서 VAE를 모델로 선정한 이유는, 음성 데이터만을 활용하여 언어를 분류해야 하는 모델의 특성상 발화자의 어투, 억양, 발음 등 음성적 특징에 대한 정교한 학습이 필요하기 때문이다. MFCC로 변환된 음성 데이터를 VAE에 적용함으로써, 단순히 특징을 벡터화하는 AE의 기능을 넘어, 생성 모델로서의 특성을 활용하여 각 레이블별로 더 정밀하게 벡터화를 수행할 수 있다고 판단하였다. 또한, 본 연구에서 사용된 데이터는 동일인의 동일 음성을 각 언어별로 녹음한 것이 아니기 때문에, 레이블 간의 일반화가 필요한 점에서 VAE가 적합하다고 보았다.

1.2.1 음성 오디오 특징 추출

VAE를 학습시키기 위해 데이터를 오디오의 특징을 추출하기에 용이하게 오디오 특징 값으로 변환하여 학습을 진행한다. 음성 특징 포착에 가장 효과적인 Mel-Frequency Cepstral Coefficient(MFCC)로 변환하여 VAE의 학습에 사용한다.

Mel은 사람의 달팽이관을 모티브로 가져온 값이다. 인간의 달팽이관에는 특수한 성질이 있는데, 저주파수 대역에서는 주파수의 변화를 잘 감지하는 것에 반해 고주파수 대역에서는 주파수의 변화를 잘 감지하지 못한다. 인간의 달팽이관 특성을 고려한 값을 Mel-scale이라고 한다.

Mel-spectrogram은 시간-주파수 도메인 표현인 spectrogram을 Mel-scale로 나타낸 것이다. Mel-spectrogram을 기반으로 추가적으로 Discrete Cosine Transform(DCT)을 적용하여 주파수 정보를

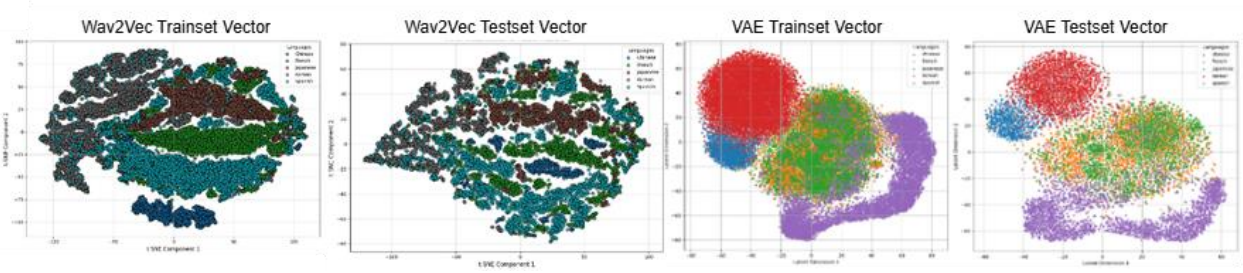


그림4. Wav2Vec 및 VAE Latent Vector 시각화

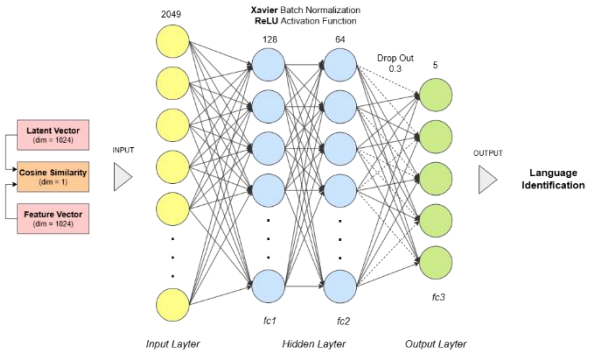


그림6. MLP Classifier 모델 구조

축소한 특징 벡터가 MFCC이다. MFCC는 발음의 음질을 더 잘 감지하기 때문에 음성 인식 분야에 훨씬 유용하다[2].

1.2.2 모델 학습 및 파라미터

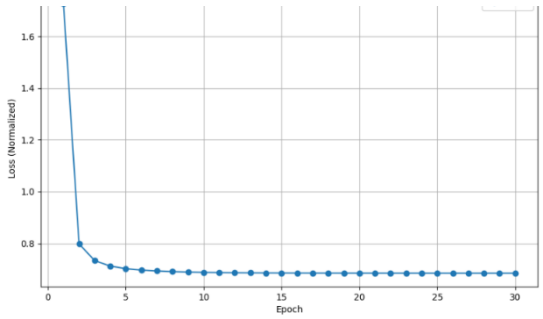
잠재 공간에서의 학습을 효과적으로 진행하기 위해, Binary Cross-Entropy(BCE)를 기반으로 한 Reconstructive Loss와 KL Divergence Loss를 결합하여 최종 손실을 정의하였다. KL Divergence Loss의 가중치(β)는 0.5로 설정하였으며, 이는 표 2에서 확인할 수 있다.

Parameters				
Epoch	30	Weight Init	He Init	
Batch Size	32	Loss weight	β	0.5
Reconstructive Loss	Binary Cross-Entropy	Divergence Loss	KL Divergence Loss	

표 1. VAE 학습 파라미터

He 가중치 초기화를 활용하여 기울기 소실 및 폭발 문제를 방지하고, 가중치를 적절히 설정하여 학습 효율성을 높였다. 학습 중 Epoch 약 20부터 손실 값의 변화가 크지 않음을 확인하였으며, 이에 따라 Batch 크기 32로 설정하고, 총 30

그림5. VAE 학습 그래프



Epoch 동안 학습을 진행하였다. 그림 5의 학습 그래프를 통해 Epoch 20 이후 Loss 값의 변화가 미미함을 명확히 확인할 수 있다.

1.3 Classifier Model

최종 언어 식별 모델은 Multi-Layer Perceptron(MLP)으로 설계되었다. 모델은 입력층, 두 개의 은닉층, 그리고 Class 개수에 맞는 출력층으로 구성되어 있다. 입력 데이터가 정형화된 고차원 벡터라는 특성과 이들 간의 비선형 관계를 학습할 필요성을 고려하여 MLP 모델을 채택하였다. 또한, Fully Connected(FC) Layer를 통해 모든 입력 차원을 학습 과정에 포함하여 복잡한 패턴을 효과적으로 학습할 수 있도록 설계하였다.

또한, Dropout 층과 Batch Normalization을 적용하여 과적합을 방지하였다. Dropout은 마지막 출력층 전에 0.3의 비율로 적용되어 특정 feature에 편향된 학습을 방지한다. Batch Normalization은 Xavier 초기화를 기반으로 가중치의 초기 분포를 조정하여 네트워크 학습 초기에 기울기 소실 및 폭발 문제를 완화한다.

그림 7에 나타난 fold별 학습 그래프를 살펴보면, fold2에서 일시적으로 손실 값이 급등하고 정확도가 낮아지는 현상이 관찰되지만, 다른 fold에서는 전반적으로 안정적인 학습 과정이 확인된다.

1.3.1 벡터 결합

Classifier의 입력으로는 총 3개의 벡터가 사용된다. 먼저 Wav2Vec을 통해 음성 데이터를 기반으로 생성된 벡터를 float32 type의 tensor로 변환하여 불러온다. 그리고, VAE를 통해 추출된 Latent Vector를 선형 레이어를 통과해 16차원을 1024차원을 매핑하는데, 이는 Wav2Vec의 feature와 차원을 동일하게 한다.

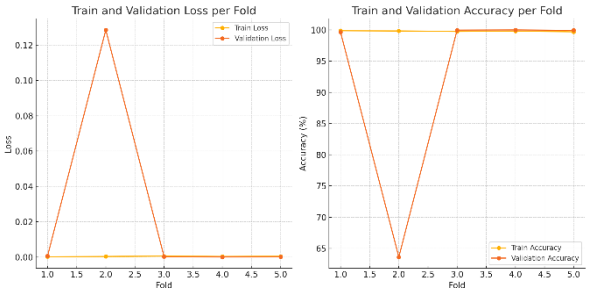


그림7. Classifier Fold별 학습 그래프

단순히 Wav2Vec과 VAE에서 추출된 Vector 뿐만 아니라 이 둘 사이의 상관성을 학습할 수 있도록 Cosine Similarity[3]를 사용해 두 벡터 간의 유사도를 측정하여 입력 값에 추가하였다. 코사인 유사도를 통해 두 Vector간의 방향적 유사성을 계산한다. 이를 통해 특징 벡터가 서로 얼마나 관련이 있는지 파악 가능하며, 관련성이 높은 경우와 낮은 경우의 패턴이 학습 가능하다.

하나의 음성 데이터에서 생성된 두 벡터는 각각 서로 다른 두 모델에서 추출된 벡터로, 두 모델의 벡터 유도 과정과 표현 방식이 상이하다. Wav2Vec 모델은 원시 음성 데이터를 그대로 입력으로 사용하여 벡터를 추출한 반면, VAE 모델은 음성 데이터를 MFCC로 변환한 뒤 이를 입력으로 사용하여 벡터를 생성한다. 이러한 차이로 인해, 두 모델에서 추출된 벡터를 단순히 결합하는 것보다, 벡터 간의 유사성이나 관계성을 도출하는 것이 더 적합하다고 판단하였다. 이를 위해 Cosine Similarity를 활용하는 방법을 도출하게 되었다.

1.3.2 모델 학습 및 파라미터

특정 레이블의 샘플 수가 다른 클래스에 비해 적을 경우, 모델은 샘플 수가 많은 레이블에 편향될 가능성이 있다. 본 연구에서 사용한 데이터셋은 레이블 간 데이터 불균형이 존재하므로, 이를 완화하기 위해 K-fold 학습과 Focal Loss를 활용하여 학습을 진행하였다.

표 2. Classifier 학습 파라미터

Parameter					
Epoch	50		Weight Init	Xavier	
Batch Size	64		optimizer	Adam	
K-folds	5		Learning rate	2e-5	
Scheduler	StepLR		Loss	Cross-Entropy	
	γ	0.5	Focal Loss	α	alter
	step	10		γ	2

K-fold 학습은 Validation 과정에서 데이터를 여러 개의 subset으로 나누어 모델을 학습하고 평가하는 방법이다. 이 방식은 과적합을 방지하는 데 효과적이며, 동일한 데이터를 5개의 동일한 fold로 나누어 각 fold를 Train set과 Validation set으로 번갈아 사용한다. 이를 통해 레이블 간의 데이터 불균형 문제를 해소하는 데 도움을 준다.

Focal Loss는 빈도가 낮은 레이블의 샘플에 더 많은 가중치를 부여하기 위해 조정 항을 사용하는 손실 함수이다. 이를 통해 모델이 이미 잘 예측하는 샘플의 기여도를 줄이고, 어려운 샘플에 대한 학습에 집중할 수 있도록 한다[4].

Focal Loss의 α는 K-fold의 각 fold 학습 시 레이블 샘플 수를 기반으로 동적으로 가중치를 설정하며, 표3을 통해 확인 가능하다. γ는 2로 고정하여 적은 샘플 수를 가진 레이블에 더욱 집중할 수 있도록 한다. Cross Entropy를 기반으로 한 손실 값은 쉬운 샘플에서는 감소하고, 어려운 샘플에서는 증가하도록 조정된다. 최종 손실 값은 각 샘플의 조정된 손실 값들의 평균으로 계산된다.

표 3. Focal Loss K-fold 별 가중치

Focal Loss α					
k-fold	Chinese	French	Japanese	Korean	Spanish
k =1~5	0.431	0.147	0.187	0.119	0.114

Schedular는 Learning Rate를 일정 epoch마다 감소시키는 역할을 한다. 학습 초기에는 높은 학습률을 사용하여 빠르게 수렴하도록 하고, 이후에는 학습률을 낮춰 세밀한 학습을 진행한다. 이 방식은 과적합을 방지하고, 손실 함수의 최저점을 안정적으로 탐색할 수 있도록 돕는다. 초기 Learning Rate는 2×10^{-5}로 설정하였으며, scheduler는 매 10 epoch마다 학습률을 변경한다. 이때 step size는 10으로 설정하고, γ=0.5로 기존 학습률의 0.5배로 감소시킨다.

Ⅲ. 데이터셋

데이터셋은 발화 데이터로 구성된 한국어, 일본어, 중국어, 스페인어, 프랑스어 데이터를 구축하였다. 언어별 데이터 양은 한국어 12,854개[5], 일본어 9,013개[6], 중국어 3,914개[7], 스페인어 14,713개[8], 프랑스어 12,061개[9]로, 언어 간 데이터 양의 편차가 큰 편이다. 모든 데이터는 Kaggle에서 제공되는 공개 데이터셋을 활용했으며, 각 언어의 발화 데이터는 특정 프롬프트에 대한 화자의 음성을 WAV 형식으로 제공한다.

표 3. 언어 별 데이터셋

Language Dataset			
	Train	Test	Total
Korean	10,283	2,571	12,854
Japanese	6,559	2,454	9,013
Chinese	2,848	1,066	3,914
Spanish	10,690	4,023	14,713
French	8,926	3,135	12,061
Total Language	39,306	13,249	52,555

Train set과 Test set은 전체 데이터의 80:20 비율로 분리하여 각각 학습 및 테스트에 사용하였다. Validation set은 Train set의 20%를 추가로 분리하여 학습 과정에서 검증 목적으로 활용하였다.

IV. 성능 평가

최종적으로 Test set에 대한 손실 값은 0.1970, 정확도는 92.85%로 모델의 성능은 전반적으로 우수한 편으로 평가되었다. 그러나 성능 개선을 위해, 예측에 실패한 데이터 중 50개를 랜덤으로 추출하여 결과를 분석하였다.

초기 모델 개발 단계에서 Focal Loss와 같은 데이터 불균형 해소 기법을 적용하기 전에는, 틀린 예측 중 True Label이 중국어(Chinese)인 경우가 90% 이상이었다. 이는 중국어 레이블 데이터 수가 현저히 부족했기 때문으로 보인다. 이를 해결하기 위해, 데이터 불균형을 완화할 수 있는 기법을 도입하였다.

```
Incorrect Predictions (Random 50 Samples):
File Name                    True Label     Predicted Label
bailen_0793.wav              spanish        french
3_2646.wav                   korean         chinese
2_0614.wav                   korean         chinese
19demarzo_1596.wav           spanish        french
bailen_2859.wav              spanish        french
meian_0104.wav               japanese       french
meian_0109.wav               japanese       french
3_2211.wav                   korean         french
batalla_arapiles_0244.wav    spanish        french
3_0723.wav                   korean         chinese
batalla_arapiles_3031.wav    spanish        french
bailen_1350.wav              spanish        french
bailen_0815.wav              spanish        french
batalla_arapiles_3806.wav    spanish        french
4_3799.wav                   korean         french
bailen_2334.wav              spanish        french
19demarzo_0691.wav           spanish        french
1_0435.wav                   korean         chinese
1_0797.wav                   korean         french
3_2600.wav                   korean         chinese
meian_4129.wav               japanese       french
meian_4525.wav               japanese       french
batalla_arapiles_2759.wav    spanish        french
19demarzo_1324.wav           spanish        french
```

그러나 그림 8에서 볼 수 있듯, 현재는 Spanish와 Korean 레이블이 틀린 예측의 대부분을 차지하고, 이들에 대한 Predicted Label로 French가 압도적으로 많이 나타나는 현상이 확인된다. 이는 데이터 불균형 해소 기법이 일부 레이블에 대해 과도하게 낮은 가중치를 부여한 결과로 추정된다. 특히, Spanish, Korean, French 레이블의 데이터 샘플 수가 상대적으로 많아, 데이터가 많은 레이블에 모델이 편향된 것으로 보인다.

하지만, 전체적인 정확도를 고려했을 때, 데이터 불균형 해소 방안을 적용한 모델이 더 우수한 성능을 보였다. 추가적으로, 데이터가 균형적으로 확보된다면, 더 나은 성능의 음성 기반 언어 인식 모델이 구현될 것으로 기대된다.

향후 상용화를 통해, 본 연구에서 제안된 모델은 다수의 사용자에게 국제적인 의사소통의 효율성을 제공하고, 언어 장벽을 극복하는 데 기여할 수 있을 것이다.

시간 관계상 Wav2Vec과 VAE에서 추출된 벡터를 각각 Classifier에 입력하거나 다양한 조합에 대한 결과를 충분히 분석하지는 못하였다. 그러나 이러한 비교 분석이 이루어졌다면 더 우수한 결과를 도출할 가능성이 있었을 것으로 판단된다.

Wav2Vec은 원시 음성 데이터를 그대로 입력받아 특징 벡터로 변환하는 장점이 있고, VAE는 MFCC를 기반으로 특징 벡터를 추출함으로써 음성 데이터를 다른 방식으로 분석할 수 있는 장점을 제공하였다. 이러한 두 모델의 특징을 결합하여 벡터를 함께 입력으로 사용한 결과가 가장 성능이 우수했을 것이라 예상된다.

그림8. Classifier 예측 실패 샘플

References

[1] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

[2] Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International journal on emerging technologies*, *1*(1), 19-22.

[3] Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information sciences*, *307*, 39-52.

[4] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., & Dokania, P. (2020). Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems, 33, 15288-15299.

[5]https://www.kaggle.com/datasets/bryanpark/korean-single-speaker-speech-dataset

[6]https://www.kaggle.com/datasets/bryanpark/japanese-single-speaker-speech-dataset

[7] https://www.kaggle.com/datasets/bryanpark/chinese-single-speaker-speech-dataset

[8] https://www.kaggle.com/datasets/bryanpark/spanish-single-speaker-speech-dataset

[9] https://www.kaggle.com/datasets/bryanpark/french-single-speaker-speech-dataset