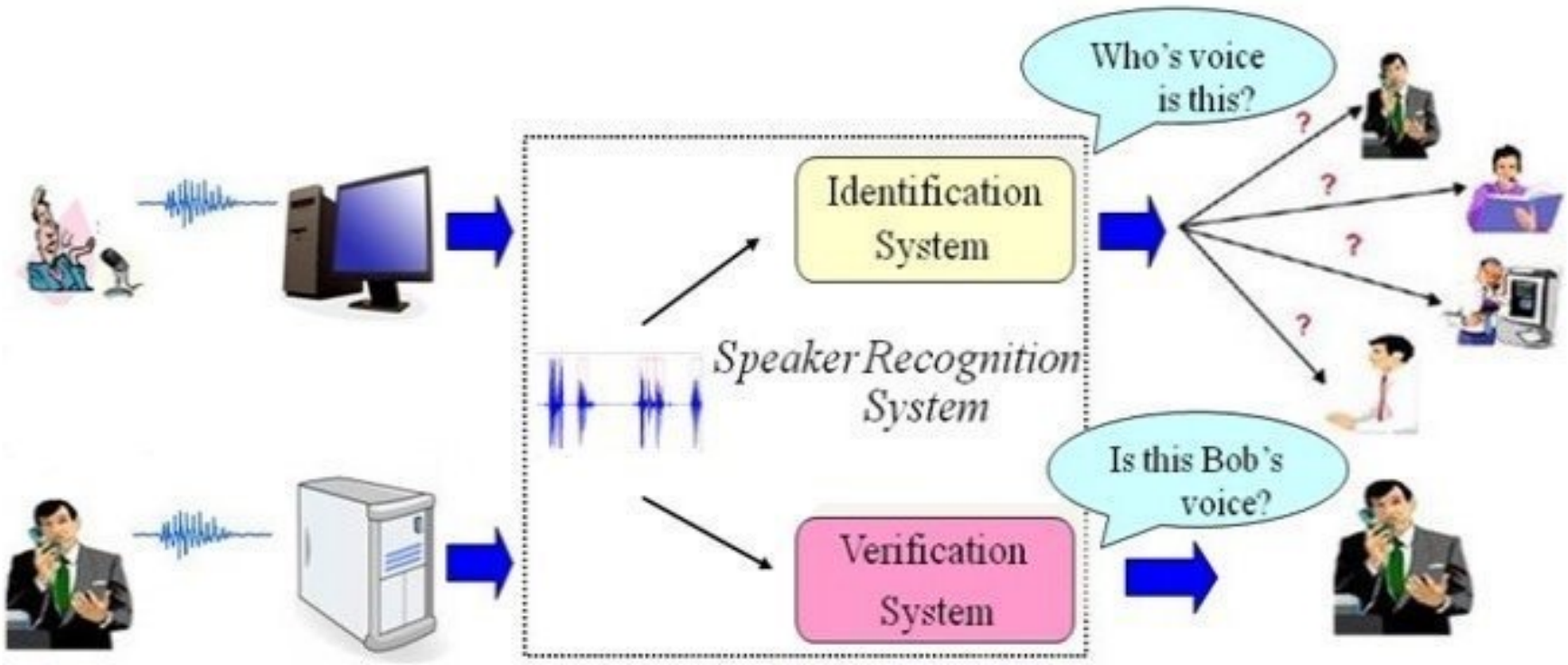
X-VECTORS: ROBUST DNN EMBEDDINGS FOR SPEAKER RECOGNITION

Speaker Identification과 Verification

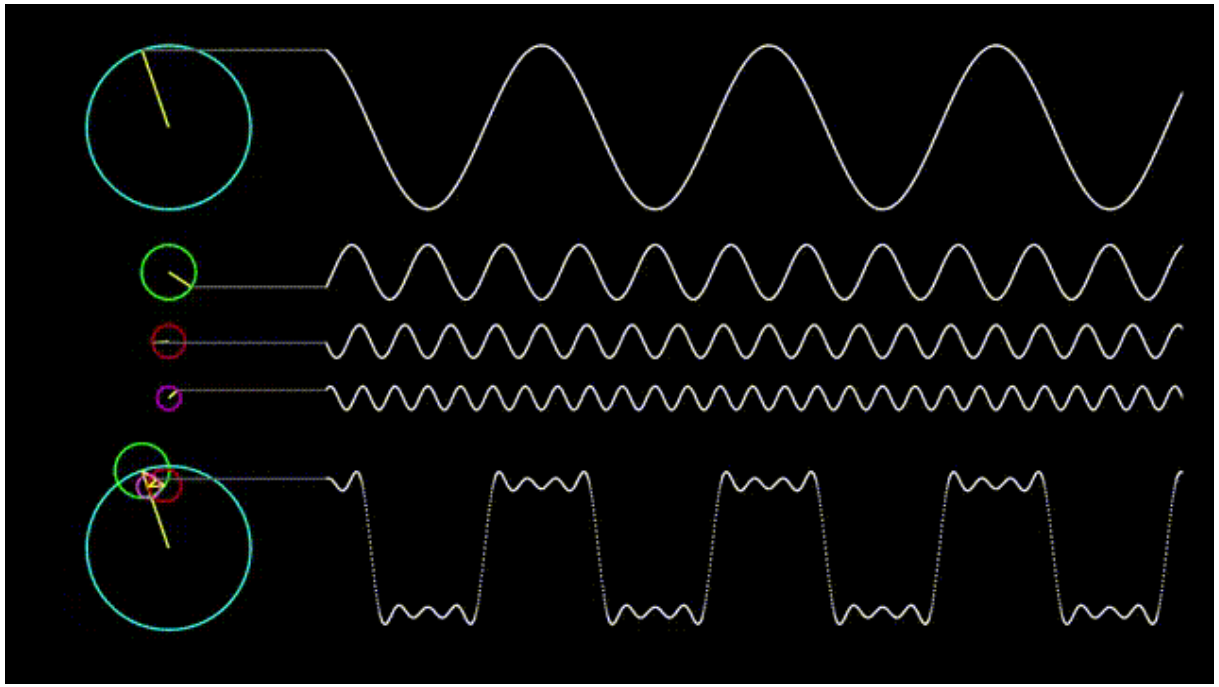


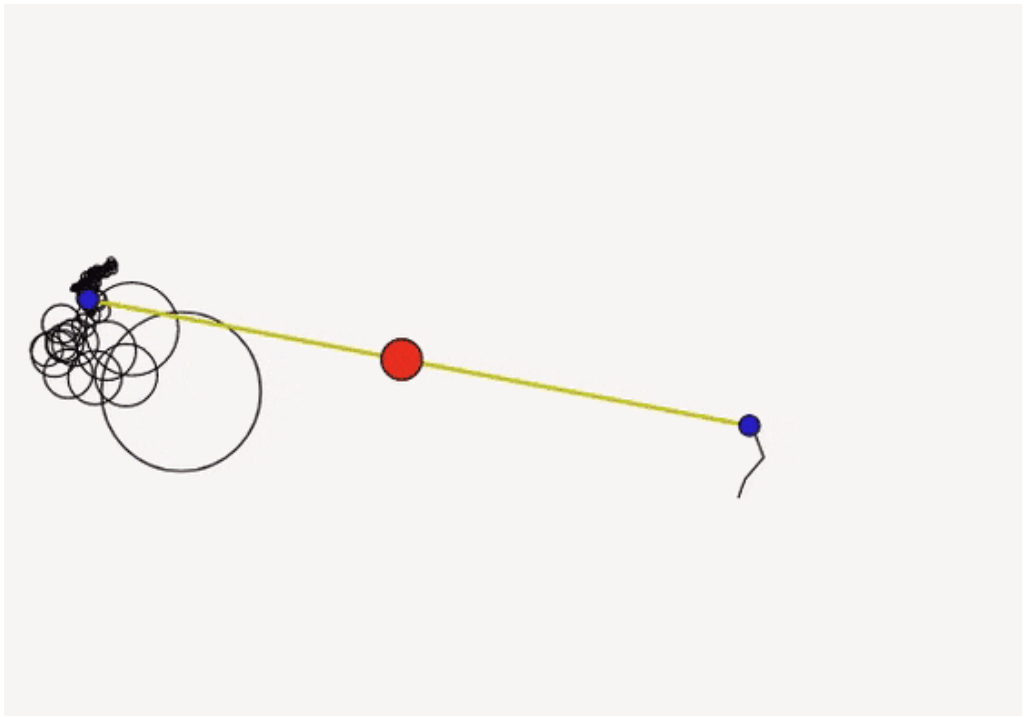
Identification : 여러 화자의 음성이 등록되었을 때 현재 들어온 음성이 누구와 가장 유사한지

Verification : 입력된 Speaker가 등록된 Speaker인지

Speech Feature Extraction

어떤 신호라도 정현파(Sinusoid)의 합으로 나타낼 수 있다.

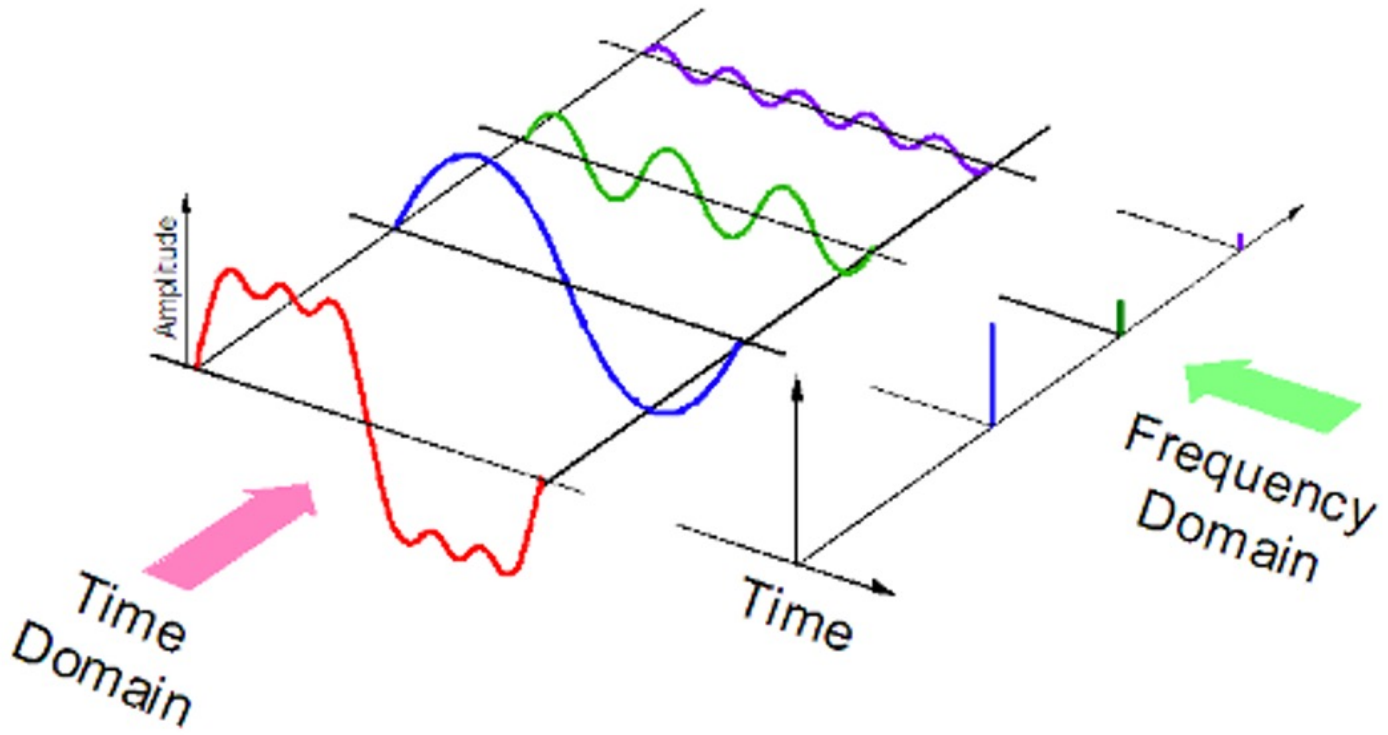




출처 : https://angeloyeo.github.io/2022/01/11/LTI_system.html

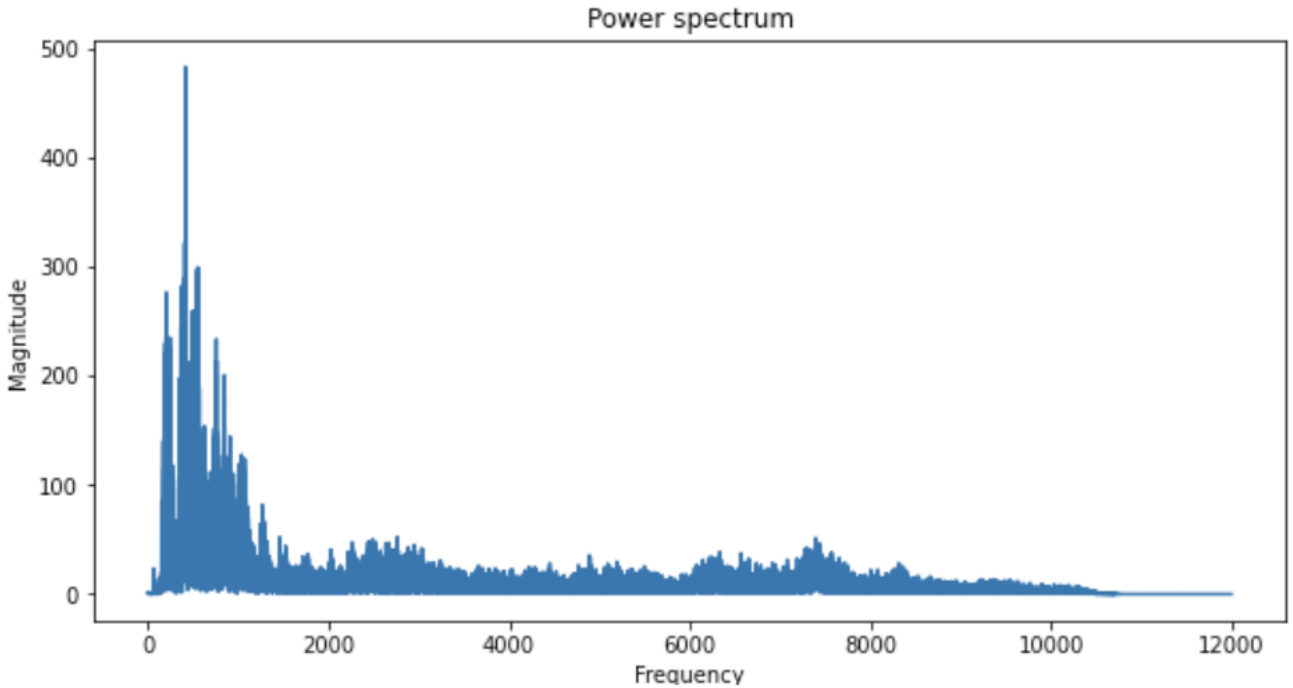
Speech Feature Extraction

Fourier Transform



Speech Feature Extraction

음성 신호에다가 FT를 바로 적용하면?



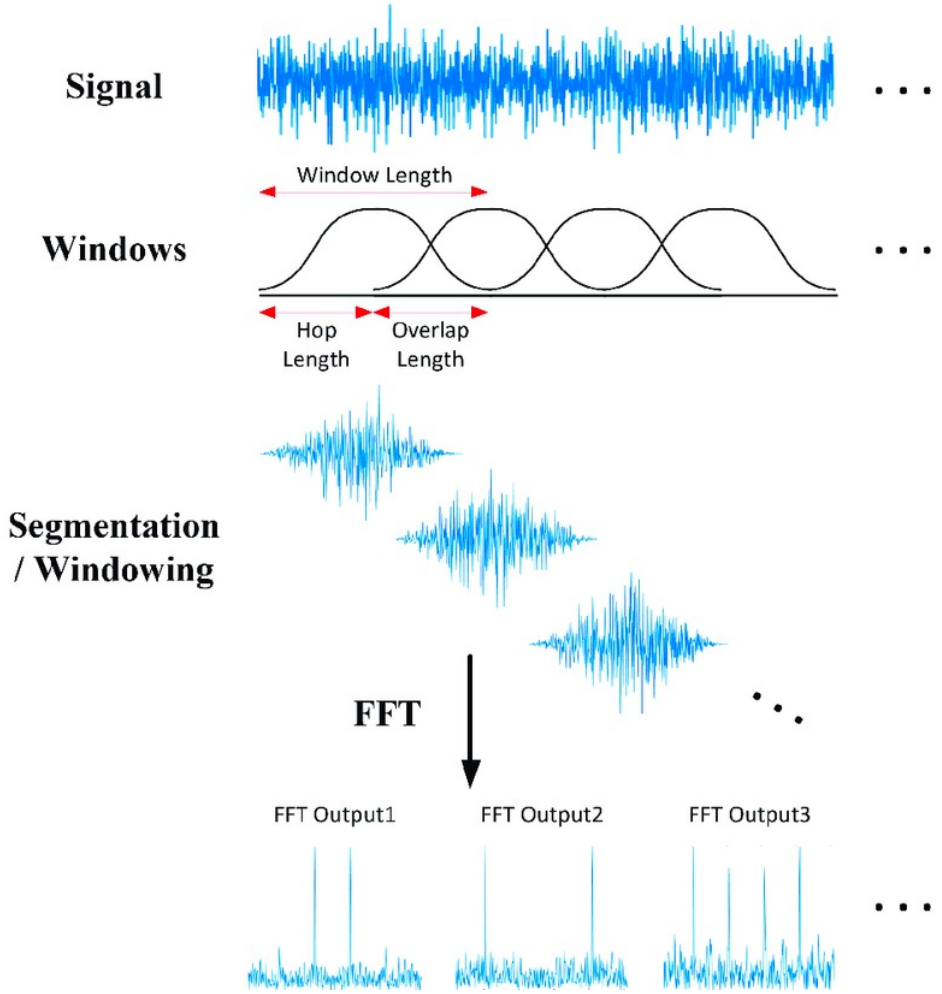
시간에 대한 정보가 모두 사라진다!

시간 정보가 사라졌다는 것은 각 주파수 성분이

언제 존재하는지 알 수 없다는 뜻

Speech Feature Extraction

Sort Term Fourier Transform(STFT)

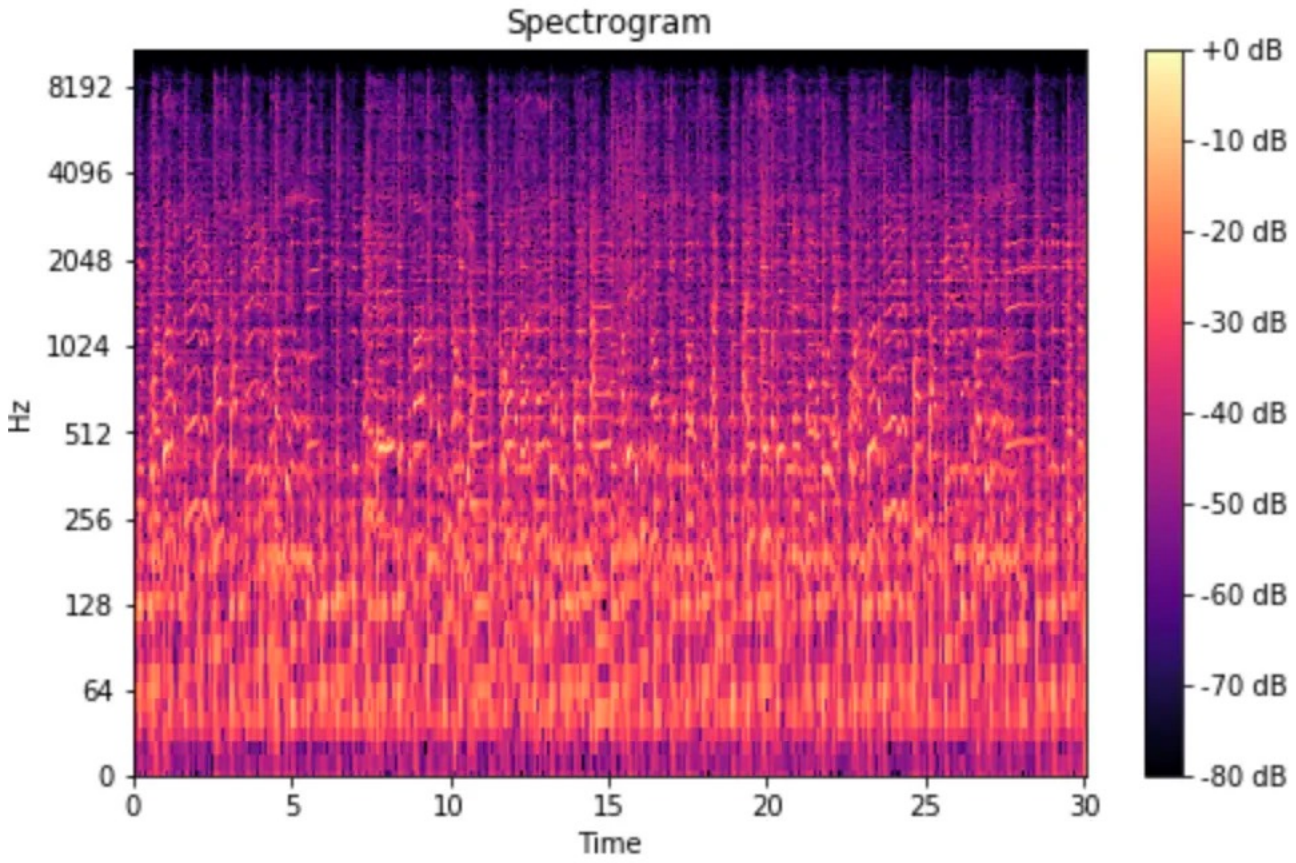


시간에 대한 정보를 위해서 signal을 일정 윈도우 만큼 이동하며 Fourier Transform을 수행한다.

출처 : https://sooftware.io/dsp_for_audio/

Speech Feature Extraction

Spectrogram



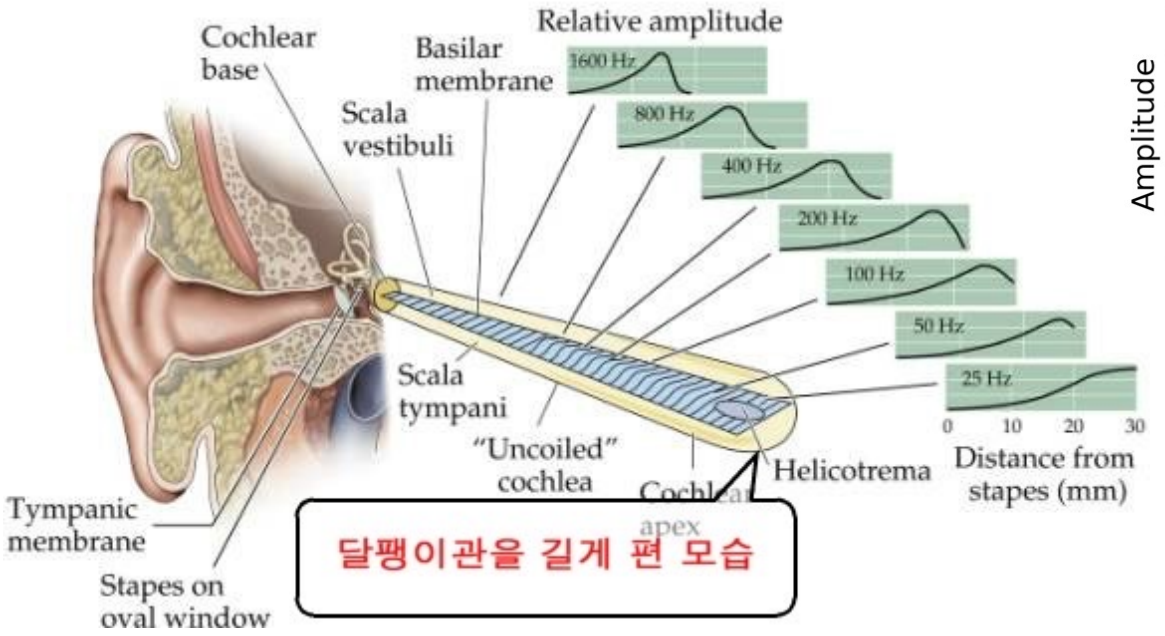
STFT를 통해 나온 결과를 스펙트로그램

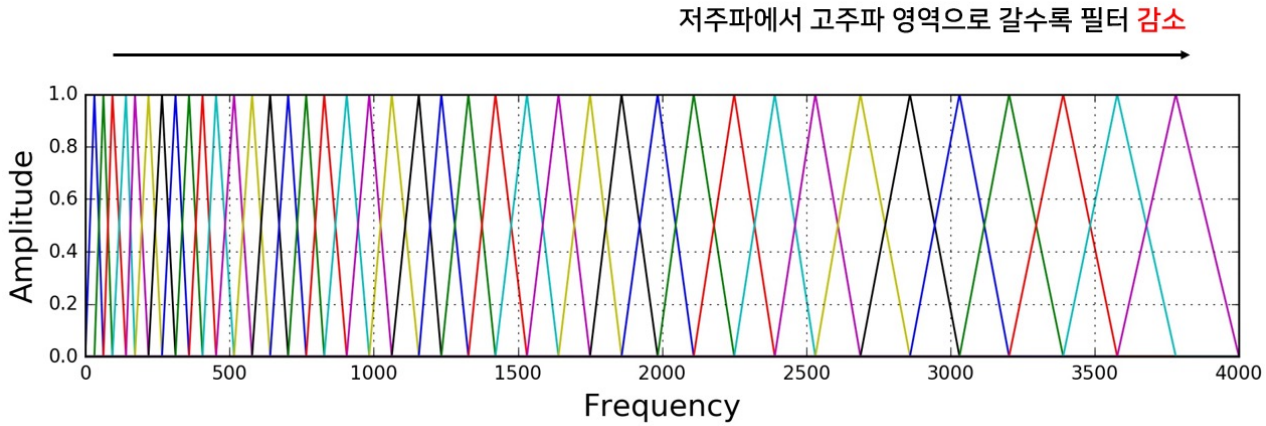
이라고 한다.

X축은 시간, Y축은 Frequency, 색은 Magnitude

Speech Feature Extraction

Mel - Spectrogram





사람의 청각기관 특성을 반영하여 사람이 인식하는

방식으로 Spectrogram을 변형

(사람의 청각기관은 저주파수 대역을 고주파수보다 민감하게 반응한다)

출처 : https://m.blog.naver.com/sooftware/221661644808

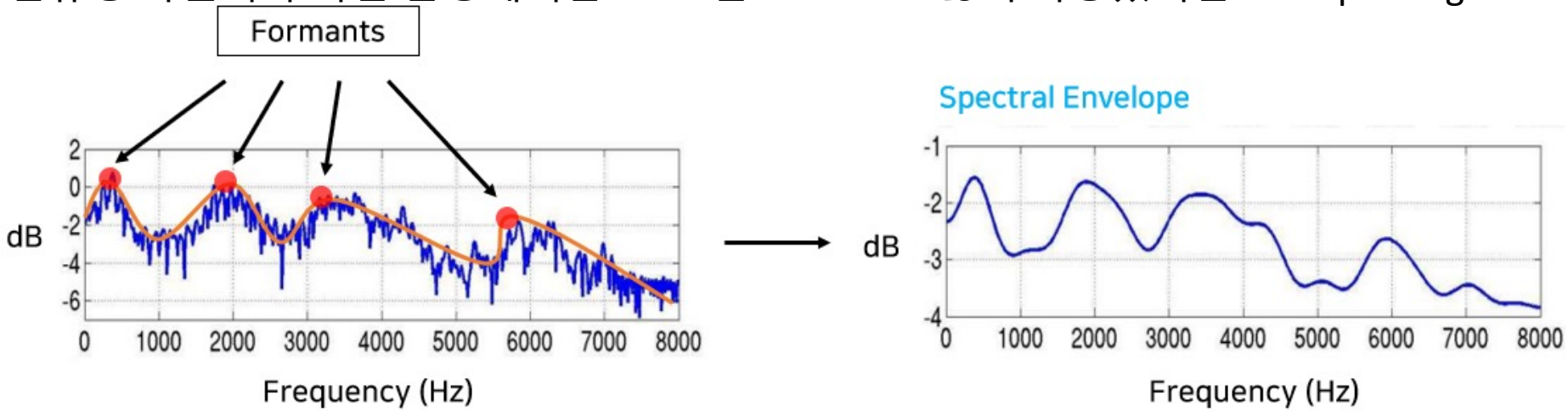
Speech Feature Extraction

MFCC - 특징벡터화

Cepstral 분석으로 지배적인 주파수를 분리하는 과정에서 MFCC 도출

Log와 Inverse FFT가 사용

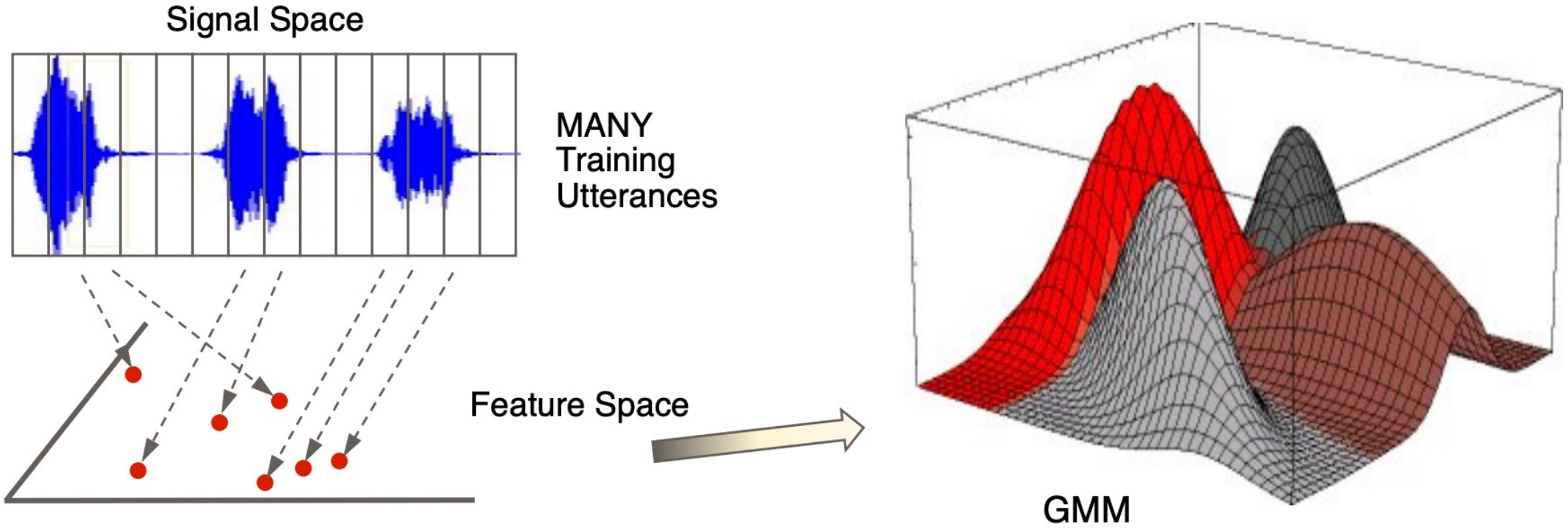
컴퓨팅 자원이 부족한 환경에서는 MFCC를 Feature로 많이 사용했지만 Mel-Spectrogram도 점점 자주 사용



출처 : https://brightwon.tistory.com/11

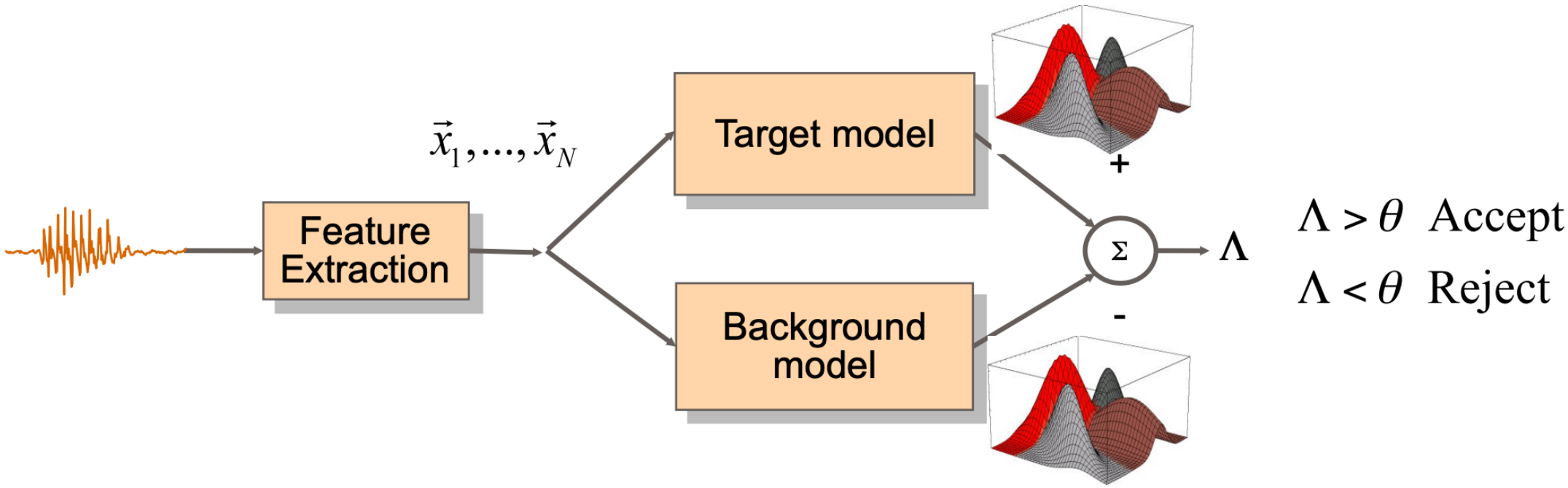
Deep Learning 이전의 기술들

초기 GMM



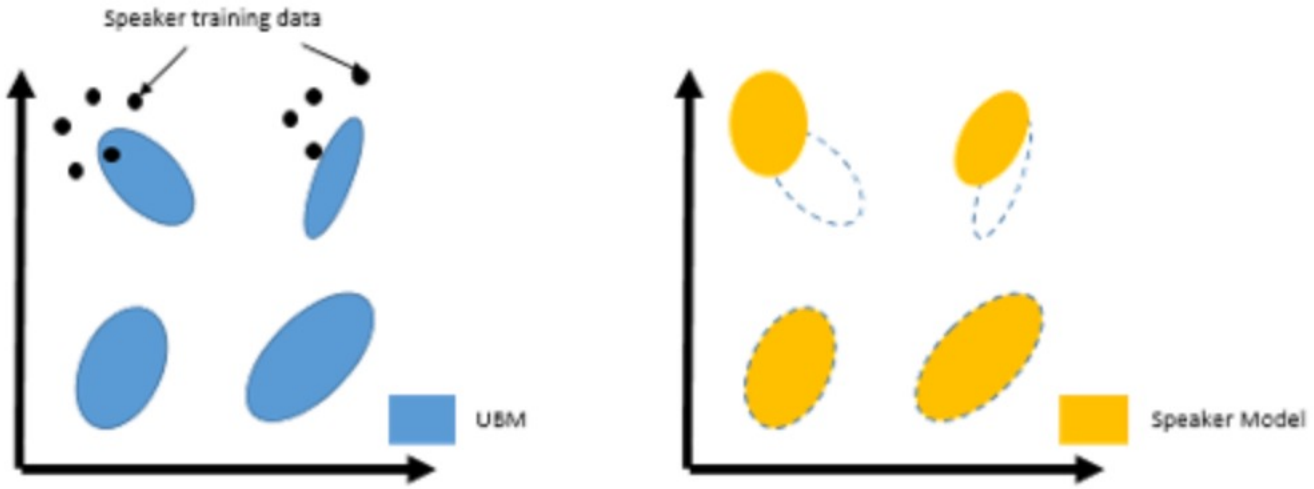
Deep Learning 이전의 기술들

GMM-UBM



Deep Learning 이전의 기술들

GMM-UBM



UBM은 Background model을 학습

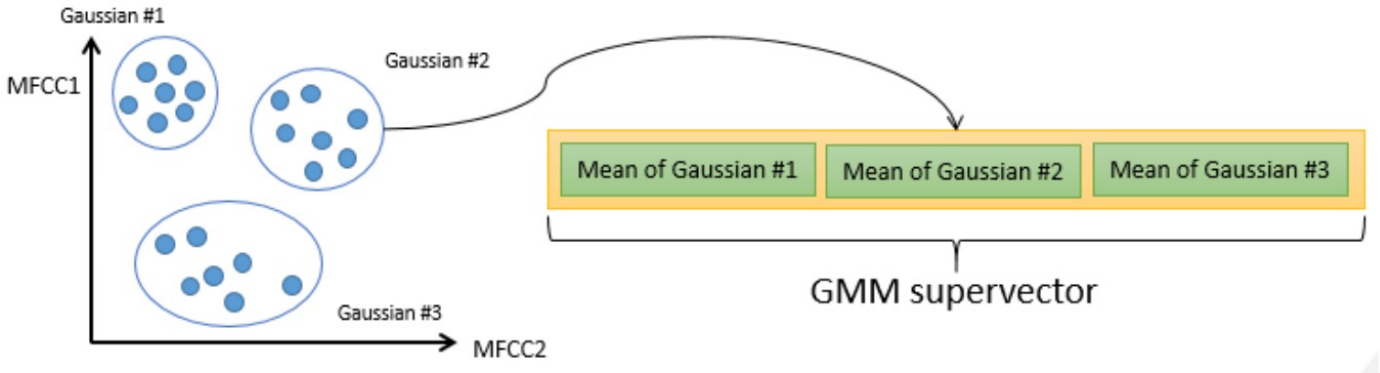
Background model은 최대한 많은 화자 및 발음 음성을 수집 후 GMM으로 clustering
→ 인간의 평균 음성 특징이 나타남

UBM의 파라미터를 초기 값으로 우리가 원하는 화자 별 GMM을 학습
→ Fine Tuning의 개념

출처 : https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=kwh1990&logNo=221485966657

Deep Learning 이전의 기술들

GMM-UBM



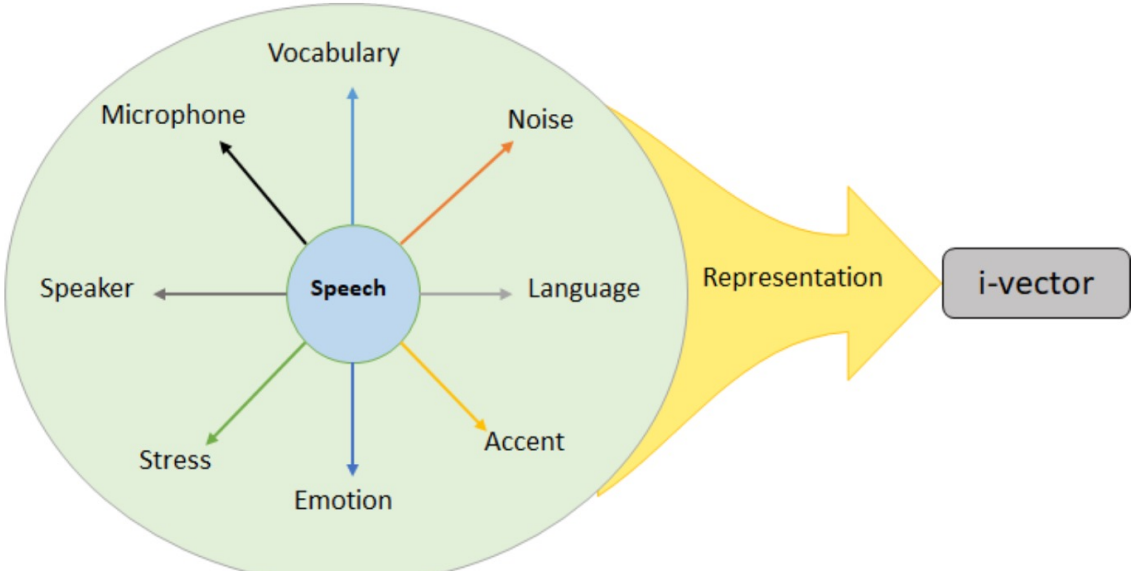
이후 각 Gaussian mixture별로 mean을 연결하여 supervector로 표현한다.

이 super vector를 SVM이나 DNN을 통해 학습시킨다.

출처 : https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=kwh1990&logNo=221485966657

Deep Learning 이전의 기술들

I - vector



$$\mathbf{m} = \mathbf{u} + \mathbf{Tw}$$

GMM supervector of the utterance

Utterance-independent GMM (UBM)

Total variability matrix (i-vector extractor)

i-vector (latent variable)

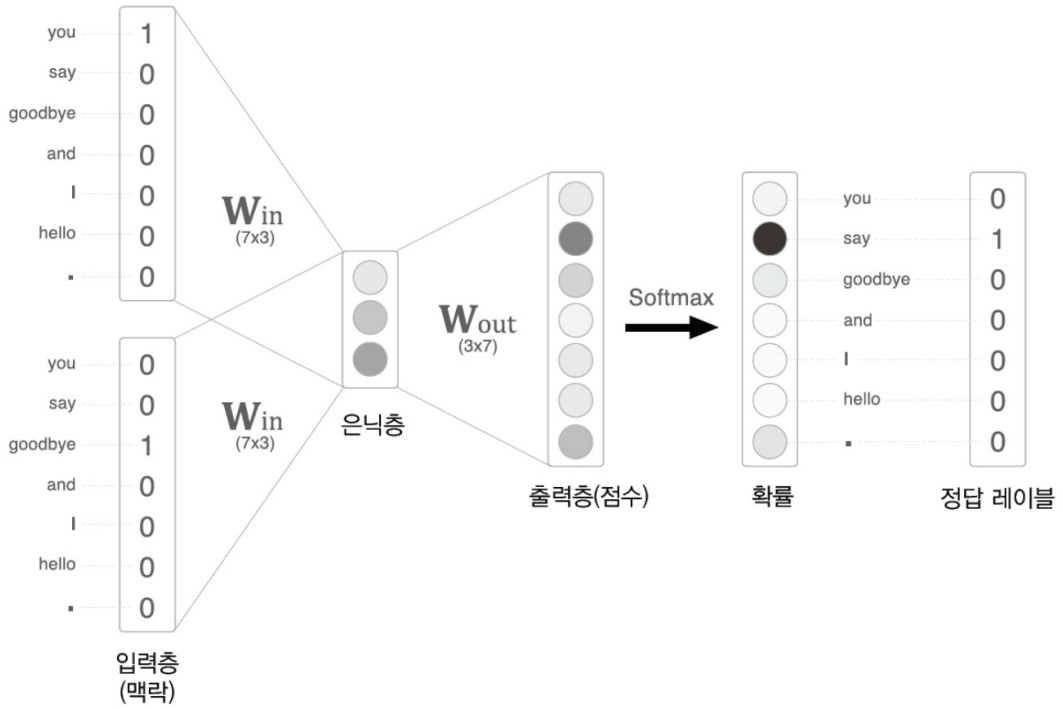
I-vector는 speaker, channel 등 모든 정보를 하나의 subspace로 표현

GMM-UBM을 기반으로 차원을 줄이고 다양한 정보를 담고있음

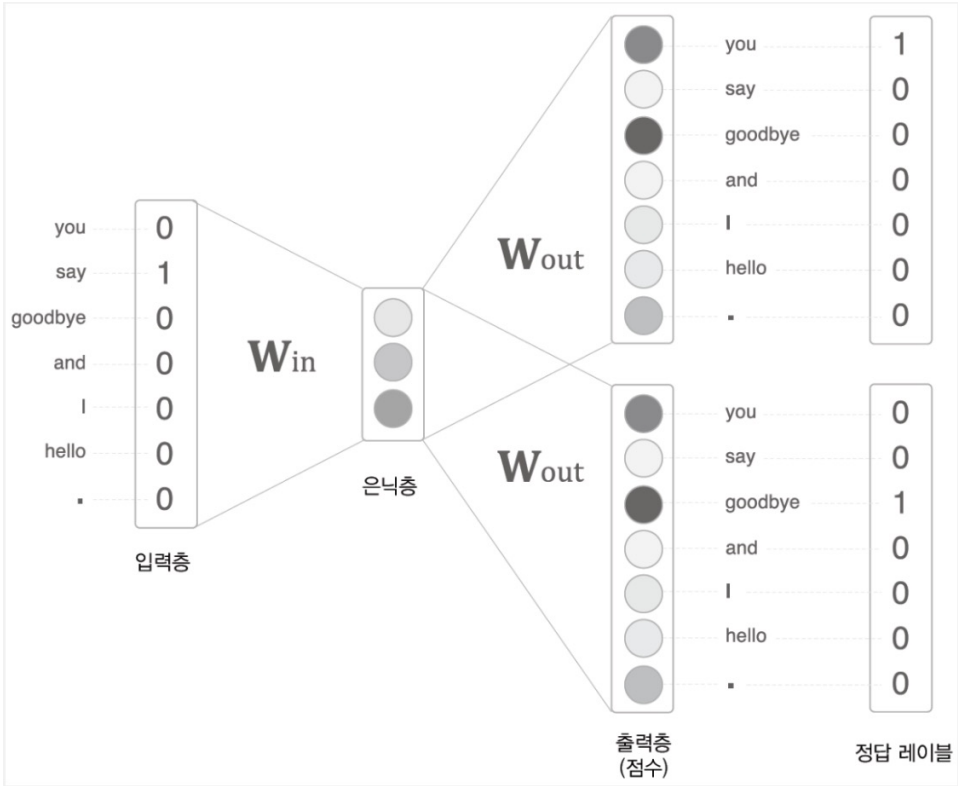
출처 : https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=kwh1990&logNo=221485966657

Deep Learning을 통한 임베딩

NLP - Word2Vec



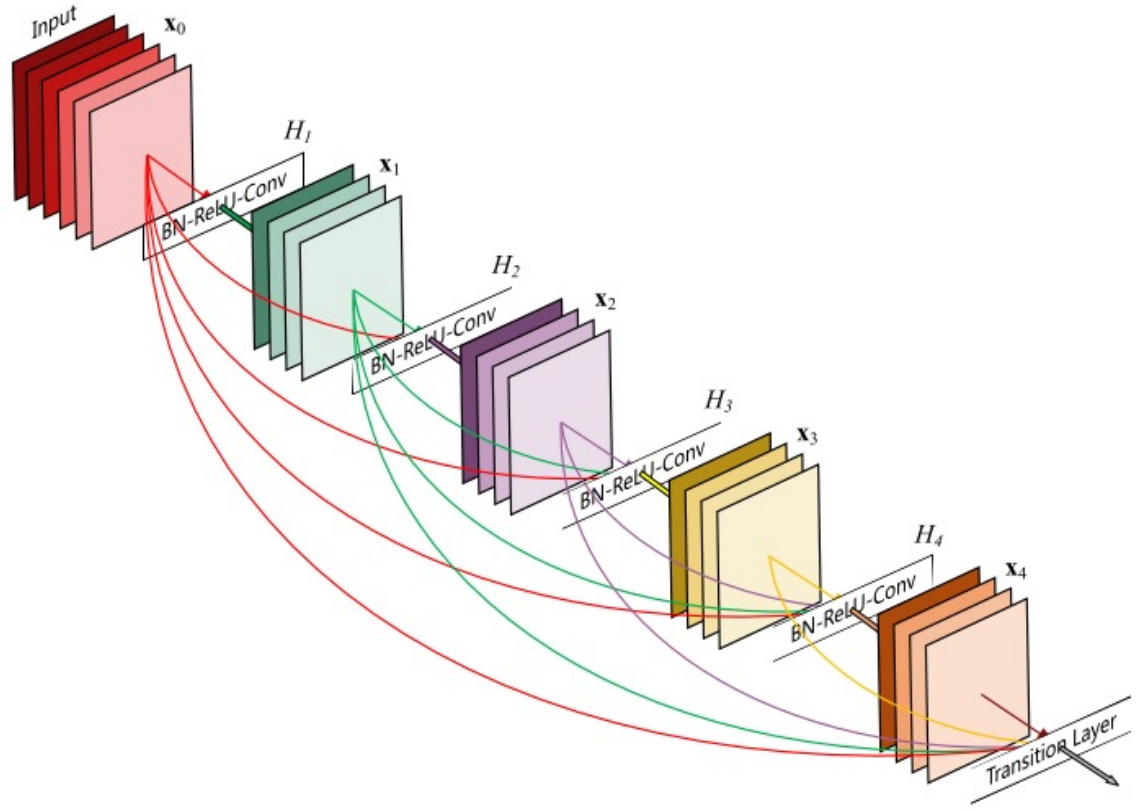
CBOW



Skip-Gram

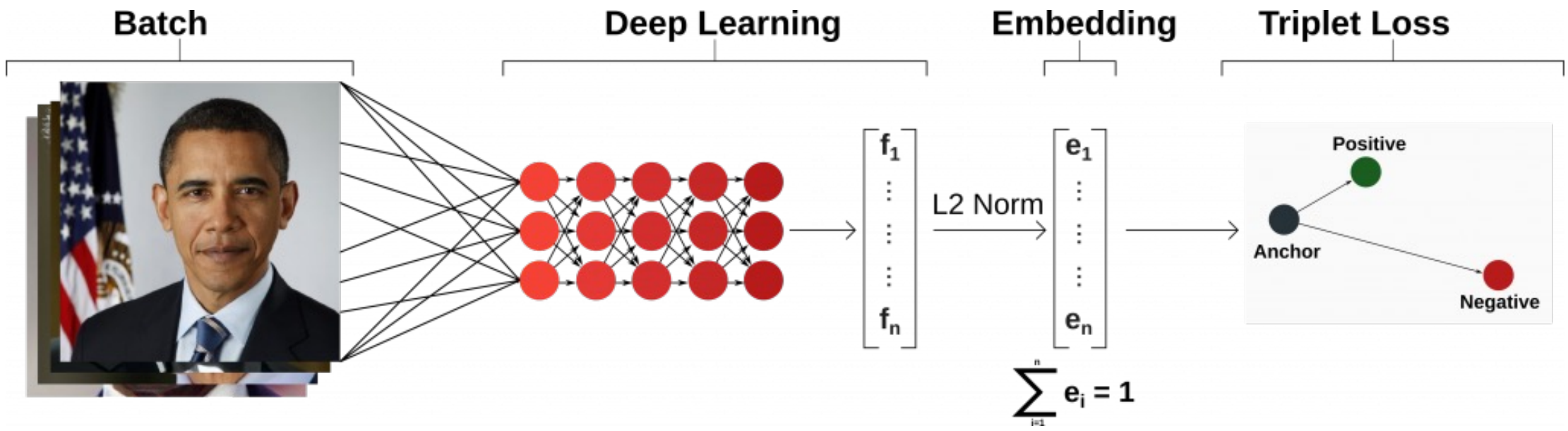
Deep Learning을 통한 임베딩

Vsion - BackBone



Deep Learning을 통한 임베딩

Vsion - FaceNet



X-vector

abstract

DNN 기반 임베딩이 large-scale training dataset이 있을 때 더 효과적이였다

그러나 대량의 labeI된 dataset을 구하는건 어려운 일이다.

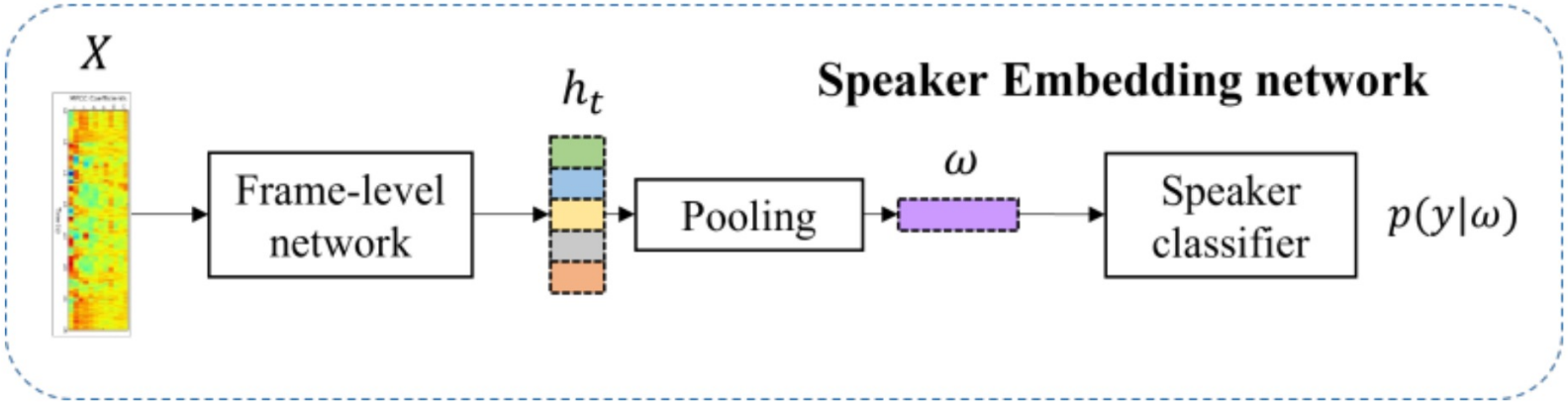
따라서 **added noise and reverberation**를 통한 augmentation으로 robustness를 올렸다.

Augmentation은 PLDA classifier에서는 효과적이였으나, i-vector에서는 아니였다.

DNN기반 X-vector에서는 supervise learning이기 때문에 augmentation의 효과를 봤다.

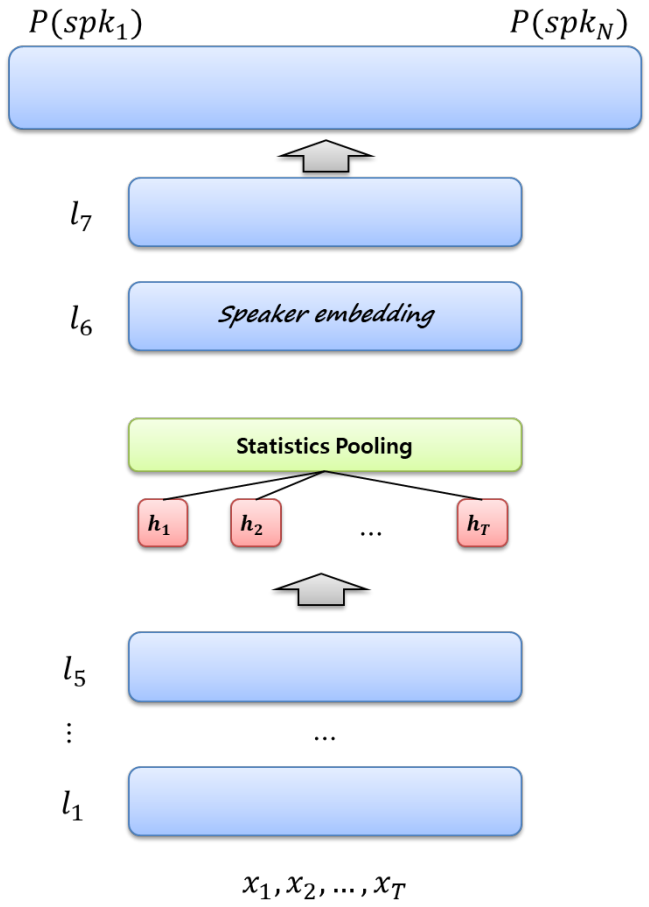
X-vector

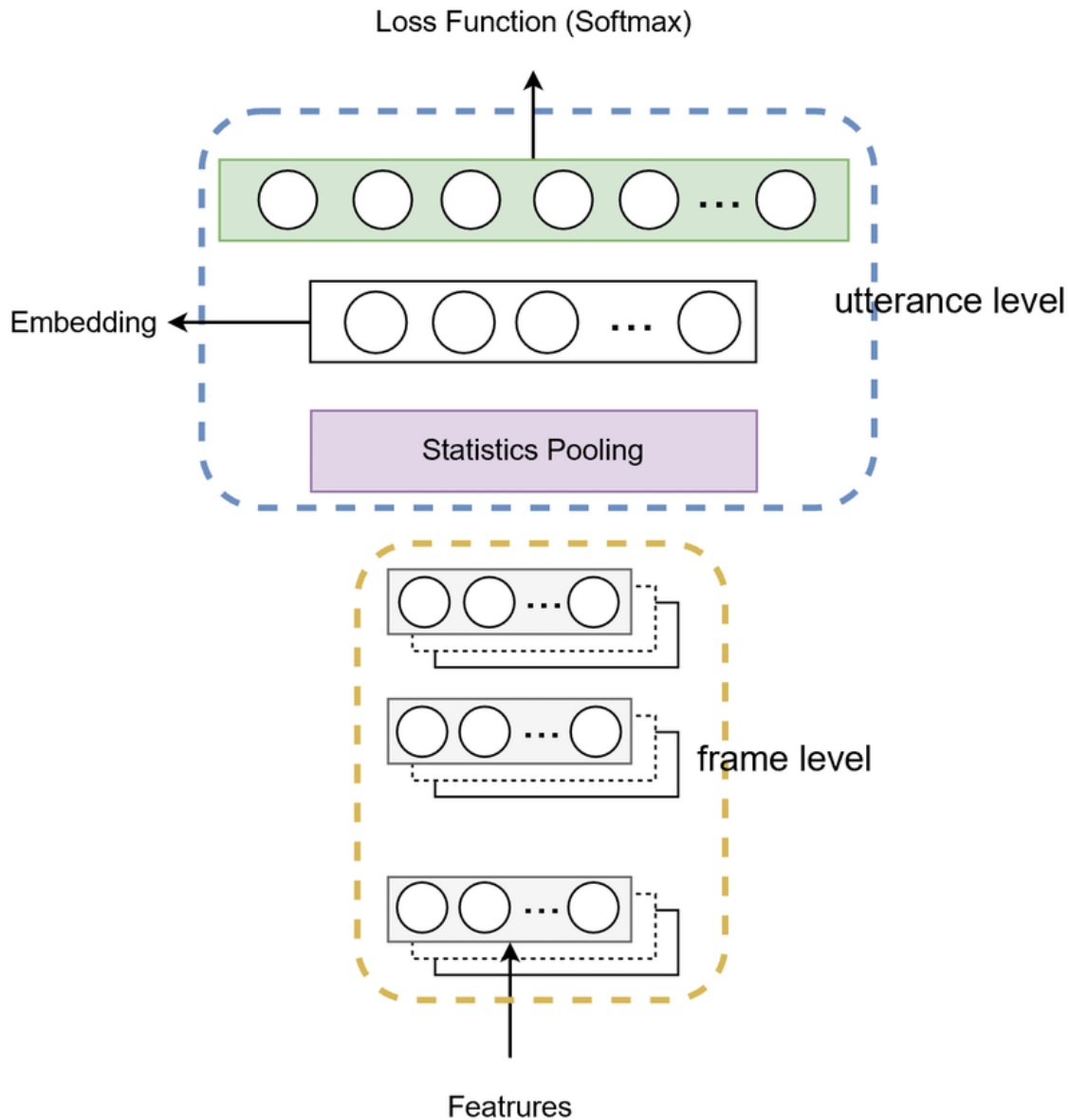
X-vector 시스템 overview

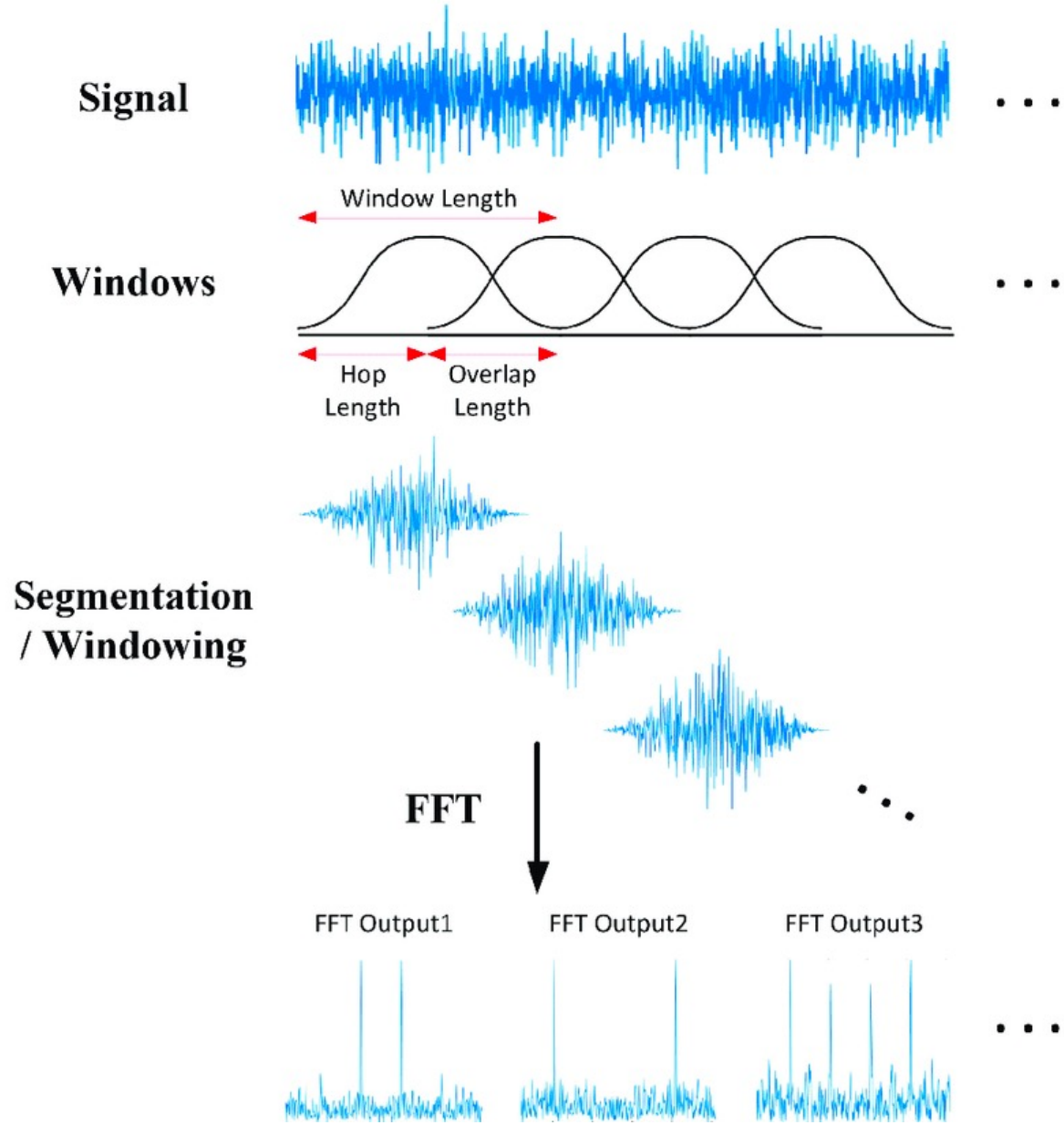


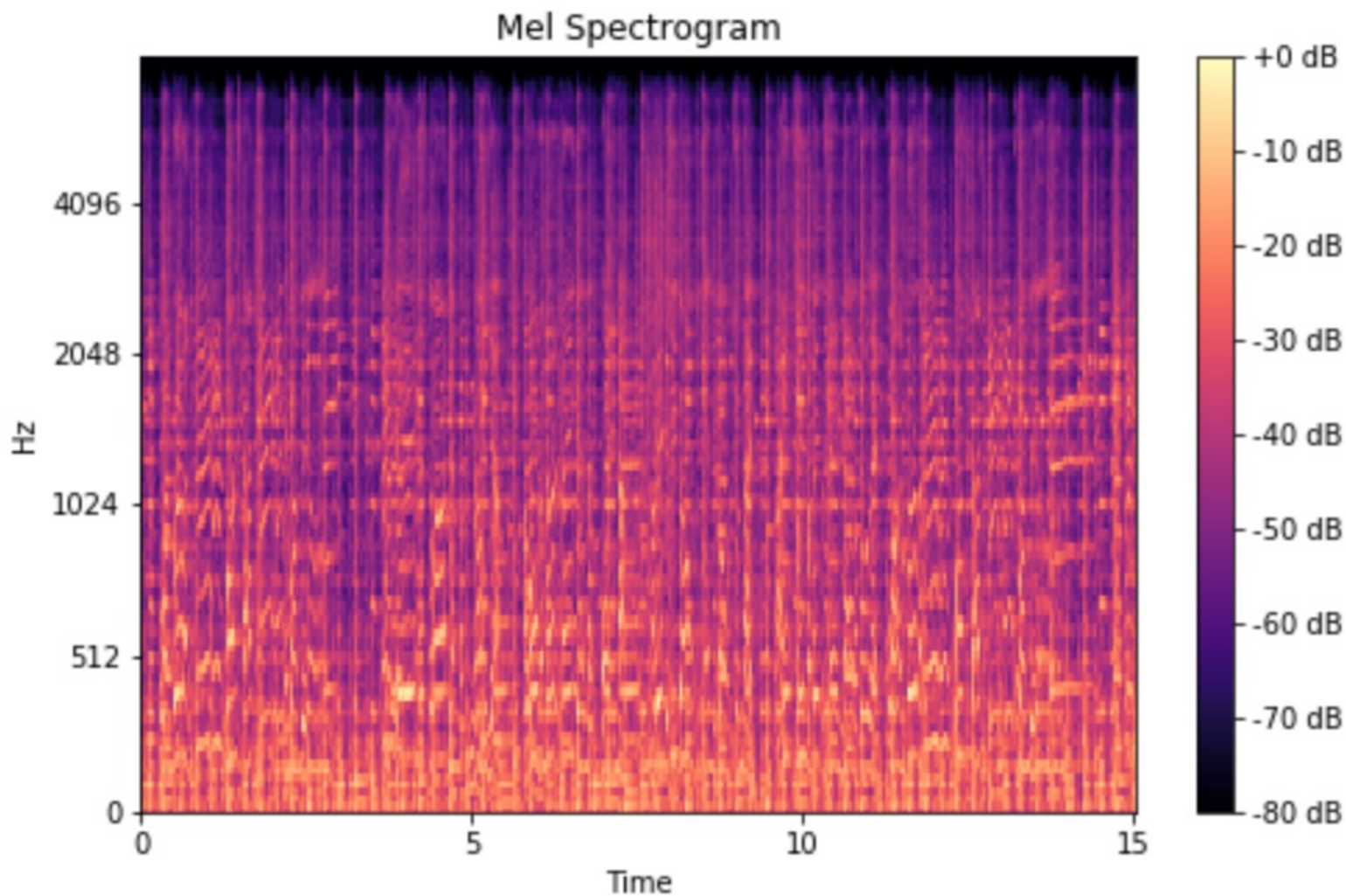
X-vector

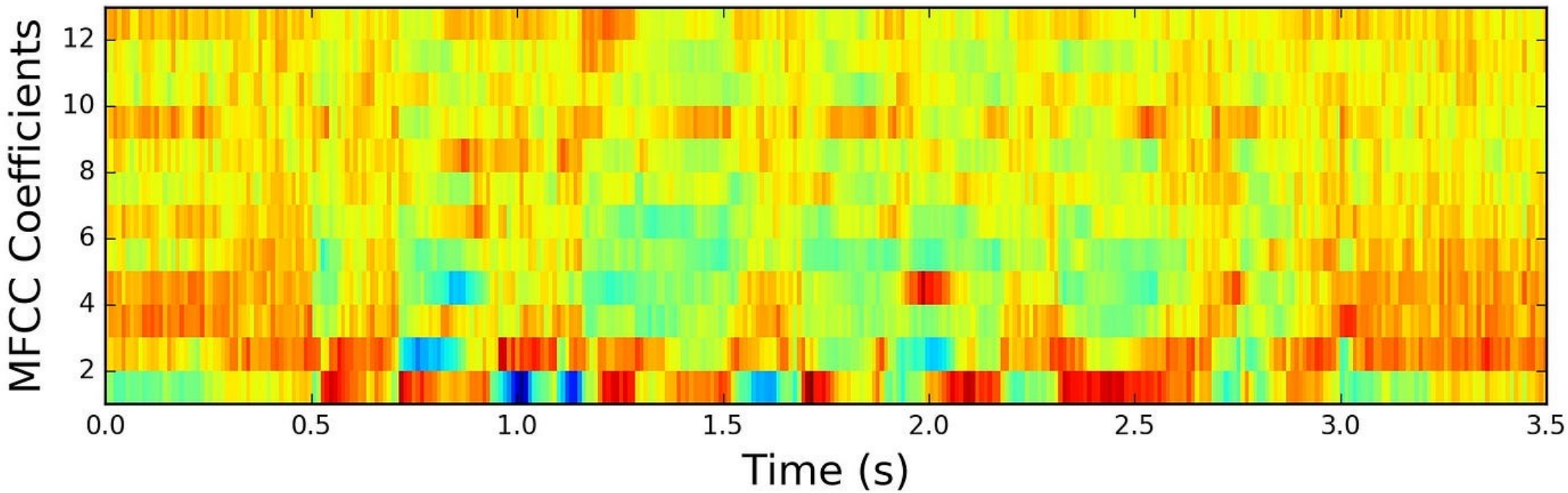
X-vector 시스템 overview




Loss Function (Softmax)
Embedding
utterance level
Statistics Pooling
frame level
Featrures


Signal
Windows
Window Length
Hop Length
Overlap Length
Segmentation / Windowing
FFT
FFT Output1
FFT Output2
FFT Output3


Mel Spectrogram
4096
2048
1024
512
0
Hz
0
5
10
15
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


MFCC Coefficients
12
10
8
6
4
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Time (s)

X-vector

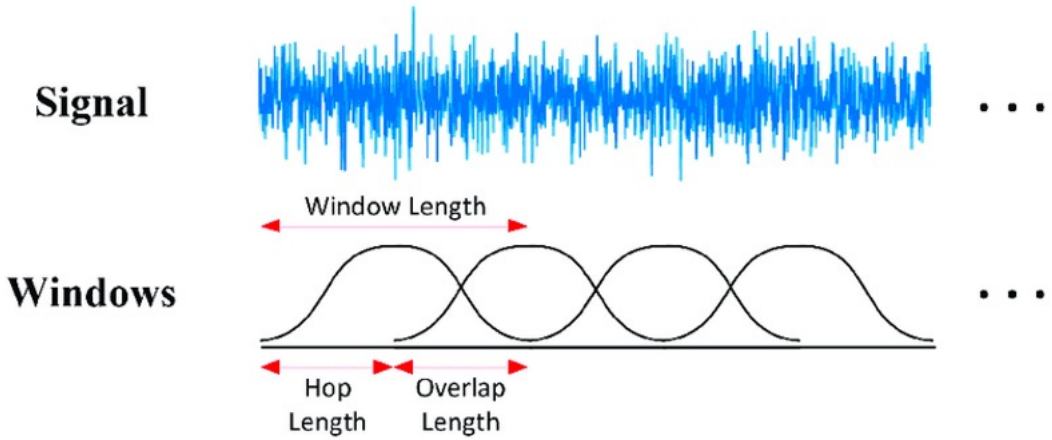
X-vector 시스템 overview

Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN

X-vector

X-vector 시스템 – 1. Filter Bank Input

Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN



1. DNN에 들어갈 acoustic feature를 추출한다.
→ Frame 단위로 MFCC같은 feature를 추출
→ VAD를 통해 speech만 이용
→ 논문에선 Frame당 24개의 feature를 사용

2. 첫 layer에는 t-2 ~ t+2까지 5개 frame을 넣는다.
→ 여기서 각 Layer는 TDNN layer를 사용한다.
(1D Conv) // 24 * 5 = 120 → 512

3. 두번째 layer는 첫번째 layer를 통과한 feature에 대하여 t-2, t, t+2 3개의 frame을 사용한다.
→ 512 * 3 = 1536

X-vector

X-vector 시스템 – 2. TDNN layer

Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN

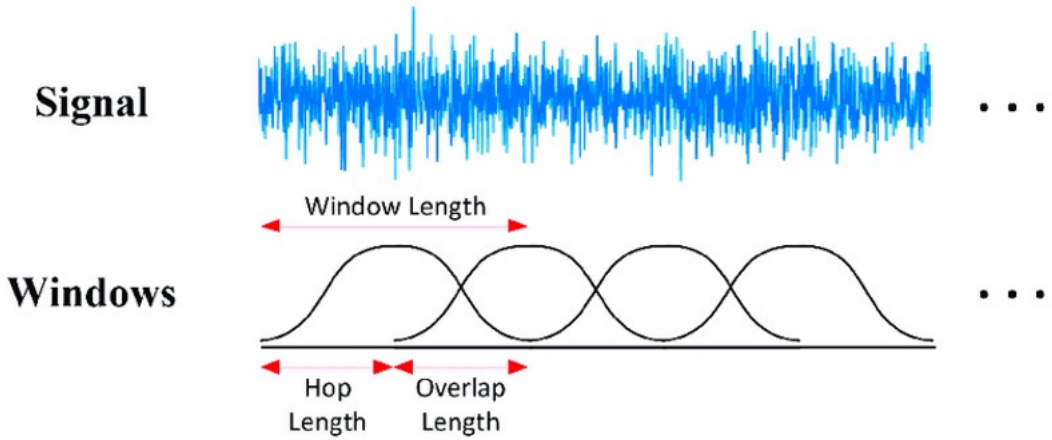
4. layer3도 마찬가지이다
→ 여기는 t-3, t, t+3 frame 사용

5. layer 4,5는 t 시점의 frame만 사용한다.

X-vector

X-vector 시스템 – 3. statistics pooling

Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	$1500T$x3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN



일반적인 pooling이 아니라 stats pooling이라는 방식을 사용한다.

입력 사이즈가 일정하지 않기 때문에 sequence의 출력을 고정된 사이즈로 바꿔주는 역할을 한다.

만약 10초짜리 음성에서 Frame이 10개가 나오고, 20초짜리 음성에서 Frame이 20개가 나왔다고 하면

mean과 std를 구한다.

Mean과 std를 concat해서 pooling 한다.

1500 개의 mean과 std --> 3000 dim

X-vector

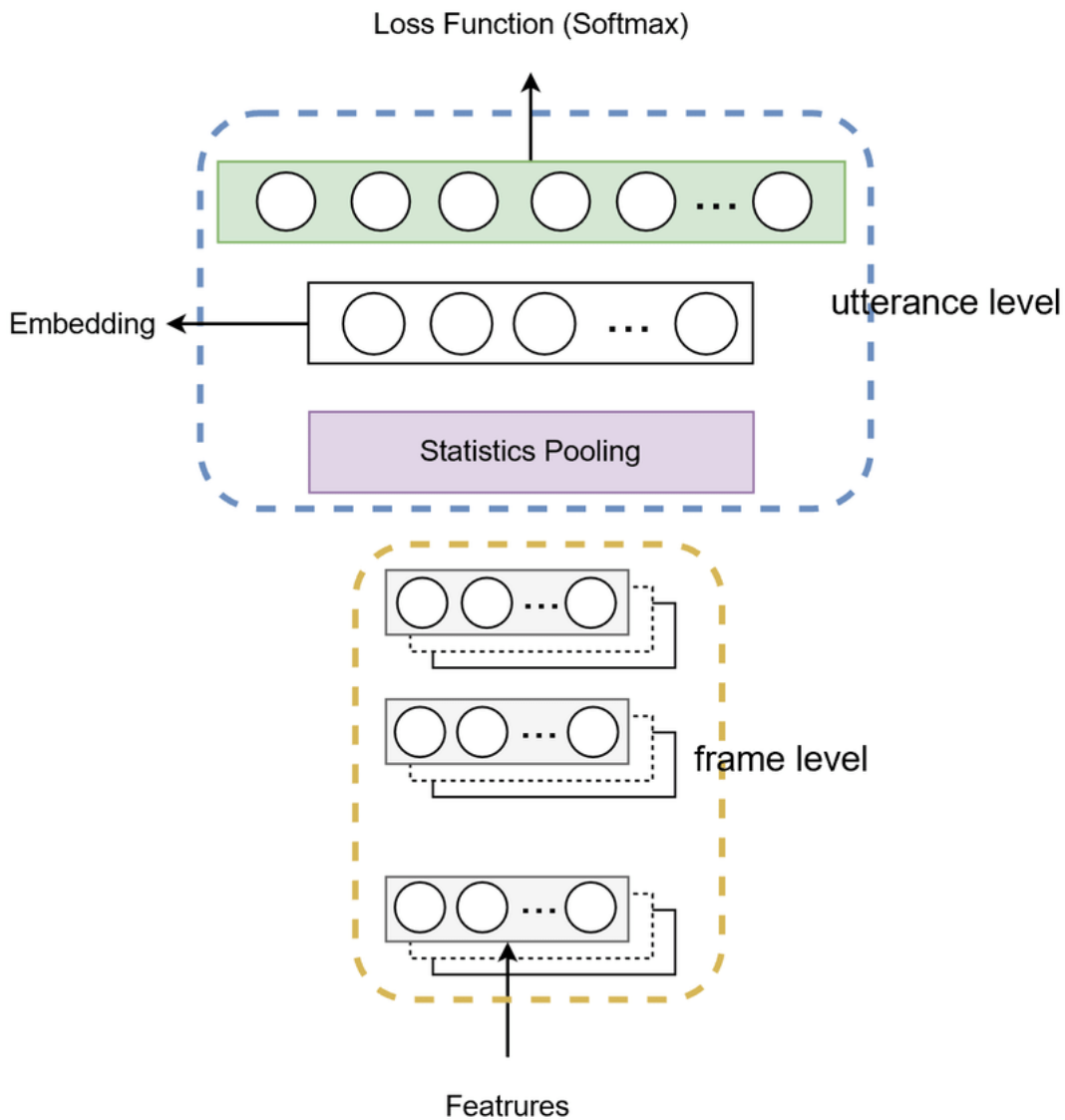
X-vector 시스템 – 4. segment layer

Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	$1500T$x3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN

이후 TDNN Layer를 차례로 통과 한 후 Softmax로 speaker classification을 진행한다.

논문에서는 1191명의 speaker identification Task를 진행

Embedding은 segment6 layer에서 추출한다.


Loss Function (Softmax)
Embedding
utterance level
Statistics Pooling
frame level
Featrures

X-vector

논문의 augmentation 방법

3.3. Data augmentation

Augmentation increases the amount and diversity of the existing training data. Our strategy employs additive noises and reverberation. Reverberation involves convolving room impulse responses (RIR) with audio. We use the simulated RIRs described by Ko et al. in [25], and the reverberation itself is performed with the multi-condition training tools in the Kaldi *ASpIRE* recipe [21]. For additive noise, we use the MUSAN dataset, which consists of over 900 noises, 42 hours of music from various genres and 60 hours of speech from twelve languages [26]. Both MUSAN and the RIR datasets are freely available from `http://www.openslr.org`.

We use a 3-fold augmentation that combines the original "clean" training list with two augmented copies. To augment a recording, we choose between one of the following randomly:

- **babble**: Three to seven speakers are randomly picked from MUSAN speech, summed together, then added to the original signal (13-20dB SNR).
- **music**: A single music file is randomly selected from MUSAN, trimmed or repeated as necessary to match duration, and added to the original signal (5-15dB SNR).
- **noise**: MUSAN noises are added at one second intervals throughout the recording (0-15dB SNR).
- **reverb**: The training recording is artificially reverberated via convolution with simulated RIRs.

X-vector

Classifier에 관한 문제점 → 논문 외 이야기

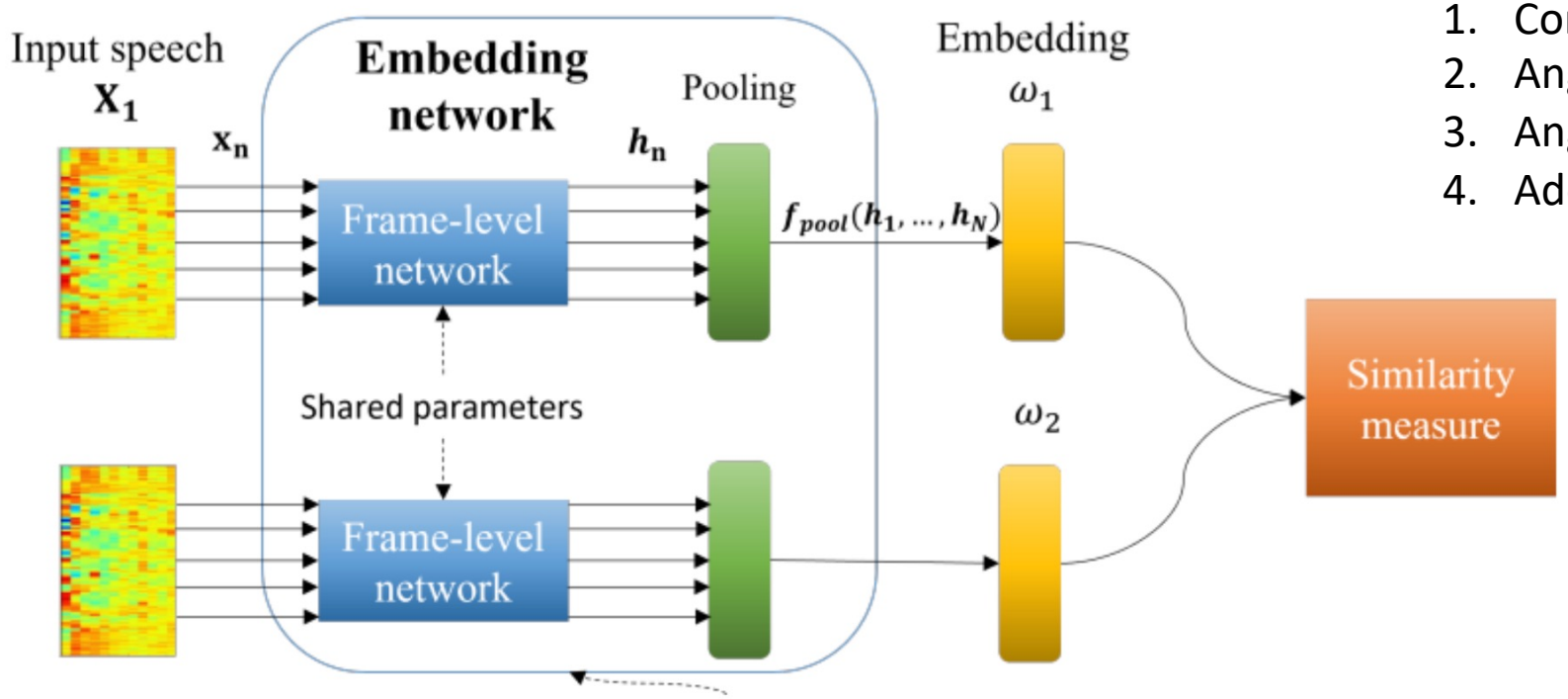
학습내에 존재하는 화자 중 입력 화자를 맞추는 Speaker Identification 같은 경우 softmax 기반 loss 학습이 적절하다.

하지만 시스템에 등록된 사용자의 목소리와 입력 목소리 둘다 학습 데이터내에 존재한다고 보장되지 않은 Verification에는 두 embedding간의 similarity를 계산한다.

softmax와 cross-entropy를 이용해도 embedding vector가 화자에 대한 정보를 최대한 추출하지만, metric mismatch는 성능을 제한할 수 있다.

X-vector

Classifier에 관한 문제점 → 논문 외 이야기



1. Contrastive loss
2. Angular softmax loss
3. Angular margin softmax loss
4. Additive angular margin softamx loss