Robust Speaker Verification Using Deep Neural Networks

ID: 45 Student: Bc. Ján Profant Supervisor: Ing. Pavel Matějka PhD.

Abstract

The objective of this work is to study state-of-the-art deep neural networks based speaker verification systems called x-vectors on wideband conditions, such as YouTube. This system takes variable length audio recording and maps it into fixed length embedding which is afterward used to represent the speaker. We compared our systems to BUT's submission to Speakers in the Wild Speaker Recognition Challenge (SITW). We observed, that when comparing single best systems, with recently published x-vectors, we were able to obtain more than 4.38 times lower Equal Error Rate on SITW core-core condition compared to SITW submission from BUT. Moreover, we find that diarization substantially reduces error rate when there are multiple speakers for SITW core-multi condition, but we could not see the same trend on NIST Speaker Recognition Evaluation 2018 Video Annotations for YouTube data.

Automatic Speaker Verification

Figure: *Speaker Verification Pipeline.*

X-vector E-TDNN Architecture

Using deep neural networks (DNN) to capture speaker characteristics is currently a very active research area. The used system is a feed-forward DNN that computes speaker embeddings from variable-length acoustic segments.

Table: *Extended TDNN x-vector architecture. x-vectors are extracted at layer 12, before the nonlinearity.*

Layer	Layer Type	Layer context	Size
1	TDNN-ReLU	[t-2,t+2]	512
2	Dense-ReLU	t	512
3	TDNN-ReLU	{t-2, t, t+2}	512
4	Dense-ReLU	t	512
5	TDNN-ReLU	{t-3, t, t+3}	512
6	Dense-ReLU	t	512
7	TDNN-ReLU	{t-4, t, t+4}	512
8	Dense-ReLU	t	512
9	Dense-ReLU	t	512
10	Dense-ReLU	t	1500
11	Pooling (mean + stddev)	Full-seq	2x1500
12	Dense(Embedding)-ReLU		512
13	Dense-ReLU		512
14	Dense-SoftMax		512

Single Speaker Recordings

Table: *Results on VAST-similar datasets without using diarization.*

System	sitwEvalC-C		voxc1	
	EER[%]	**$DCF^{min}_{0.01}$**	**EER[%]**	**$DCF^{min}_{0.01}$**
BUT i-vector [Novotný et al., 2016]	9.34	0.713		
x-vector tel	7.16	0.559	9.00	0.676
E-TDNN tel	5.90	0.519	7.74	0.599
E-TDNN 16k	2.60	0.242	2.77	0.286
E-TDNN 16k HT-PLDA	**2.13**	**0.221**	2.73	0.304
x-vector 16k LC	2.74	0.268	2.99	0.330

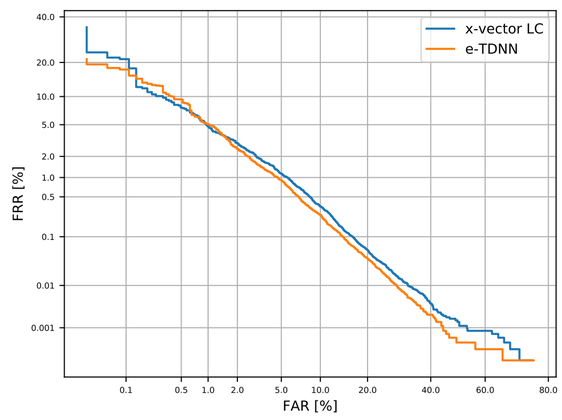

Figure: *Detection error tradeoff curve for systems trained on 16k Hz VoxCeleb1 and VoxCeleb2 data for sitwEvalC-C condition.*

Multi Speaker Recordings

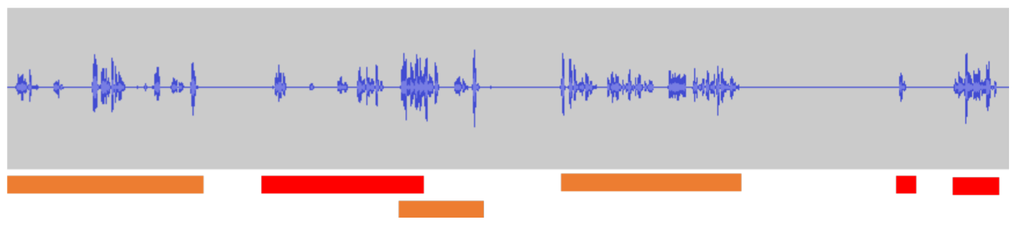

Figure: *Example output of diarization on single channel audio. Different colors at the bottom indicate different speakers.*

Table: *Results for domain specific systems on VAST-similar datasets.*

System	Diarization	sitwEvalM-C		sre18EvalVAST	
		EER[%]	**$DCF^{min}_{0.01}$**	**EER[%]**	**$DCF^{min}_{0.01}$**
E-TDNN 16k	no	5.09	0.338	13.33	0.758
E-TDNN 16k	yes	4.02	0.269	12.35	0.738
E-TDNN 16k	iterative clustering	2.87	0.262	12.03	0.789
x-vector 16k LC	yes	4.14	0.292	13.59	0.713