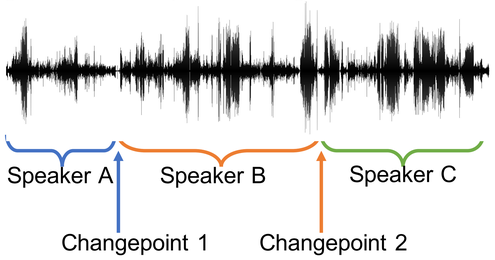
Pre-Training of Speaker Embeddings for Low-Latency Speaker Change Detection in Broadcast News

Leda Sari[1], Samuel Thomas[2], Mark Hasegawa-Johnson[1], Michael Picheny[2]
[1] Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
[2] IBM Research AI

1. Speaker Change Detection

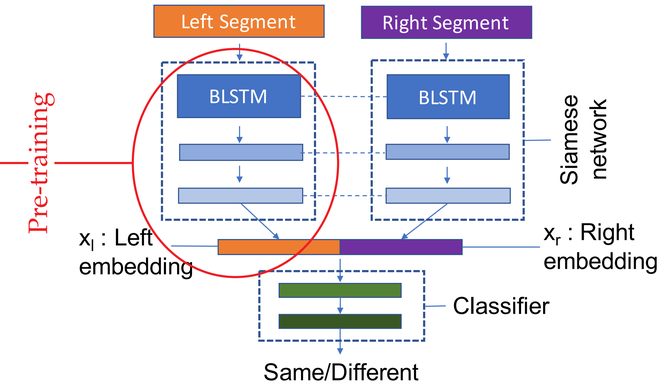
- We are interested in the time points at which the change happen
- Speaker identity is not important



- Existing methods are based on comparing the features from two consecutive segments
- For an online application, **can we operate with segments of at most 2 seconds?**

2. Method

1. Gender classification
2. Contrastive loss training
3. Triplet loss training



$$\mathcal{L}_c = \sum_{m=1}^{M} \delta[s_l^{(m)} = s_r^{(m)}] d(x_l^{(m)}, x_r^{(m)}) + \delta[s_l^{(m)} \neq s_r^{(m)}] \max(0, \Delta_c - d(x_l^{(m)}, x_r^{(m)})) \quad (1)$$

$$\mathcal{L}_{tri} = \sum_{m=1}^{M} \max(0, \Delta_{tri} + d(x_a^{(m)}, x_p^{(m)}) - d(x_a^{(m)}, x_n^{(m)})) \quad (2)$$

3. Results: Accuracy

Dataset:

- 144 hours of audio from LDC HUB4 Broadcast News
- Training segments have duration of 2s
- Sampled 500k pairs, 329 triplets for training

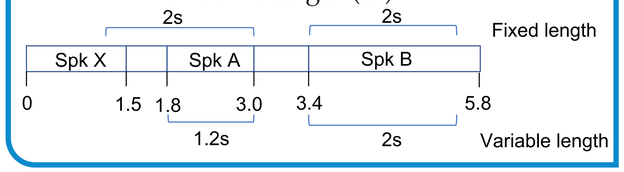
Feat.	Net.	Pre-train	Freeze Embed.	Accu.
PLP	C	-	-	52.2
i-vector	C	-	-	86.6
PLP	S+C	Gender	Yes	76.9
PLP	S+C	Gender	No	78.1
PLP	S+C	Contrast	Yes	77.4
PLP	S+C	Contrast	No	87.5
PLP	S+C	Triplet	Yes	82.7
PLP	S+C	Triplet	No	**89.0**

Results:

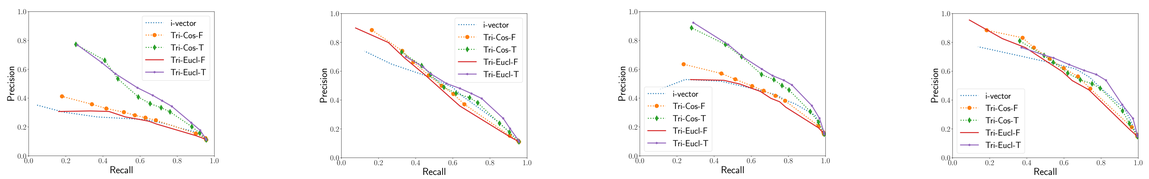
- Among three pre-training methods **triplet loss** is the best
- Using Euclidean distance is slightly better than the cosine distance

4. Test Setup

- 10 audio files are chosen for test
- Left-right comparisons are performed around segment boundaries rather than using sliding windows for **low-latency**
- Choice of segments
  - Based on segment type
    1. ASR
    2. Ground truth
  - Based on segment duration
    1. Variable length ($\leq$ 2s)
    2. Fixed length (2s)



5. Results: Precision-Recall and F-measure



(a) Variable length-ASR (b) 2s-ASR (c) Variable length-Ground truth (d) 2s-Ground truth

	ASR boundary		Ground truth boundary	
	Variable	2-second	Variable	2-second
i-vector	0.3150	0.4902	0.5036	0.6109
Tri-Eucl-F	0.3332	0.4591	0.4722	0.5736
Tri-Eucl-T	**0.4746**	**0.5323**	**0.6141**	**0.6511**

- Relative improvements in F-measures as compared to i-vectors are
  - 50.7% in the highly mismatched condition (ASR-Variable length)
  - 6.6% in the matched condition (Ground truth-2s)
- Score combination of i-vector and Triplet-T system performs 5% better on 2s segments

6. Conclusions

1. Jointly trained Siamese network and the classifier performs better than classifying i-vectors
2. Siamese embeddings are more robust to the duration mismatch between training and test segments
3. Siamese embeddings perform better than i-vectors for $\leq$2s segments which is important for achieving low-latency