Leveraging LSTM Models for Overlap Detection in Multi-Party Meetings

Neeraj Sajjan†, Shobhana Ganesh†, Neeraj Sharma†, Sriram Ganapathy†, Neville Ryant‡

neerajsajjan.ec13@rvce.edu.in, shobhana224@gmail.com,neerajww@gmail.com,sriramg@iisc.ac.in,nryant@gmail.com

†Learning and Extraction of Acoustic Patterns (LEAP) Lab, Indian Institute of Science, India
‡Linguistic Data Consortium, University of Pennsylvania, USA

Why is overlap detection challenging?

Overlap segments comprise of speech from more than one talker (shown in Fig. 3).

- Occur in multi-talker conversational speech setting, such as meetings, debates, and broadcast news.
- Usually small in duration, have speech and non-speech overlaps, overlapping talkers energy ratio is time-varying, and the far-field recordings contain reverberation and ambient noise.
- Acoustic features characterizing overlap segments are not well defined.

Accurate overlap detection can improve analysis of conversational speech recordings.

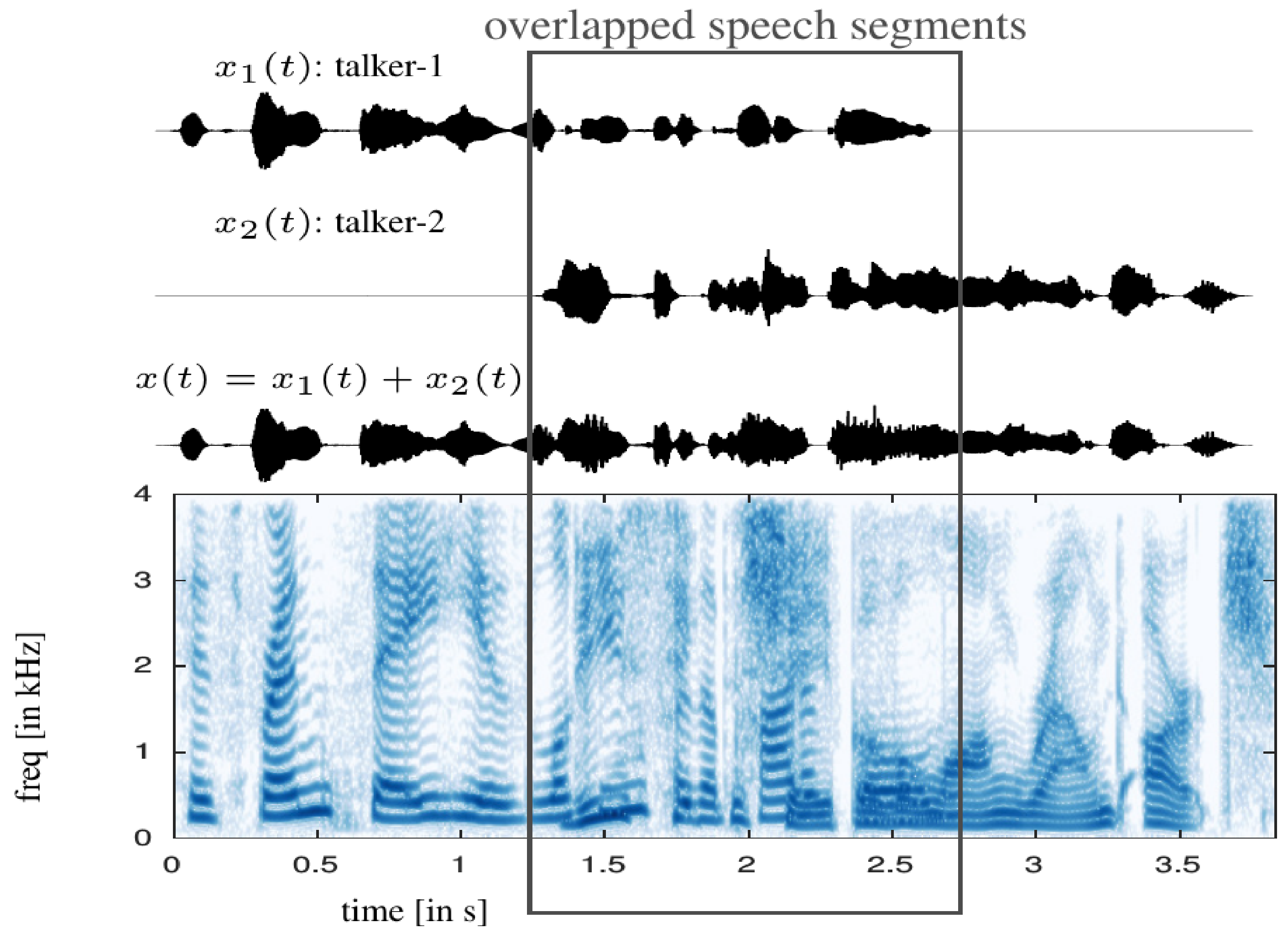

Figure 1: Synthetically overlapped speech segments.

Proposed idea

▷ Feature designing for overlap detection is complicated. To circumvent this we propose using:

- time-frequency representations, namely, mel-spectrograms and gammatone spectrograms as features
- a context of 100 msec to obtain decision on every 10 msec short-time segment
- deep learning architectures for classification.

▷ To train and test the approaches we use: *(i)* synthetically designed overlaps using TIMIT dataset, *(ii)* AMI corpus, and *(iii)* forced aligned AMI corpus.

Prior art

- make use of handcrafted features such as sample kurtosis (kurt), spectral-flatness measure (SFM), MFCCs, and harmonicity, etc.
- model single and overlap speech classes using GMMs
- perform poorly on conversational speech recordings

Table 1: Detection accuracy % with GMMs (Baseline approach) [1].

Dataset	Features	Single	Overlap	Avg.
TIMIT	kurt.+SFM+MFCC+D	59.6	69.4	64.5
AMI	kurt.+SFM+MFCC+D	43.1	61.9	52.5

Our Contribution

The previous work on overlap detection used guassian mixture models (GMMs) with a variety of handcrafted features.

We show that,

- instead of designing specialized features, spectrograms can be used,
- recurrent networks models like LSTM are effective in overlap detection.

We perform evaluation on,

- synthetically designed overlap speech dataset built from the TIMIT dataset, and meeting conversations from the augmented meeting interaction (AMI) corpus.

Results show that,

- proposed approach perform better than existing methods,
- usefulness extends to meeting conversations,
- Viterbi decoding with the posteriors from the network boost accuracy.
- improved overlap detection benefits speaker diarization.

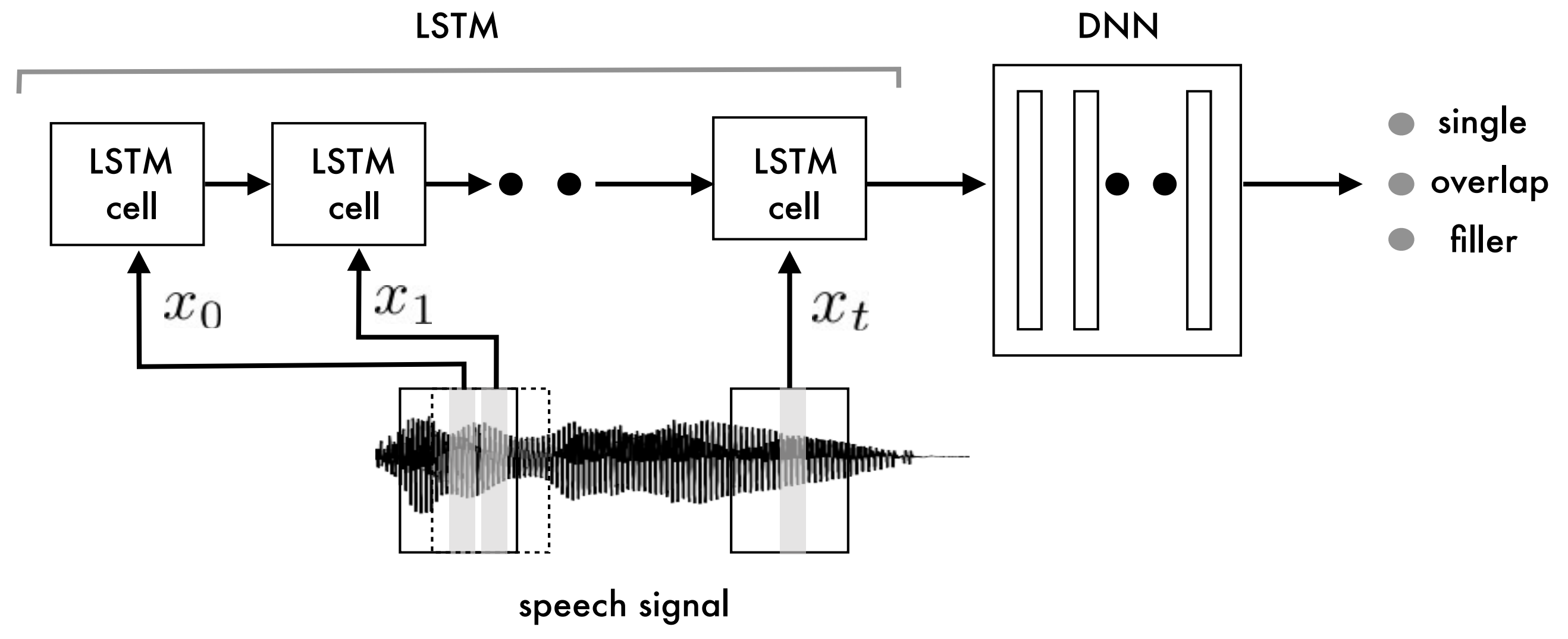

Figure 2: Proposed LSTM architecture for overlap detection.

Evaluation

- Label the dataset into 3 classes - single speaker, overlap and filler
- Experiment with DNNs, CNNs, and LSTMs.
- Use data augmentation for AMI corpus using synthesized TIMIT overlaps, in training.

Results

Table 2: Detection accuracy % with fbank features

Model	TIMIT Dataset			AMI Dataset		
	Single	Overlap	Avg.	Single	Overlap	Avg.
DNN[3 layers]	73.0	87.0	**79.9**	56.3	73.0	64.7
lstm[512 cells]	73.7	83.1	78.4	76.0	60.6	**68.4**
blstm[256 cells]	78.7	79.5	78.9	51.4	75.3	63.4
blstm[512 cells]	72.5	87.0	79.7	58.3	71.8	65.1
Conv [1 layer]-lstm[512 cells]	89.8	52.0	71.8	49.5	74.5	62.0
Conv [3 layers]-lstm[512 cells]	87.0	63.0	74.9	57.8	68.0	63.0

Table 3: Detection accuracy % on AMI Meeting Dataset using different features with the LSTM model.

Features	Single	Overlap	Avg.
gammatone	66.3	75.1	70.7
gammatone + kurt.+SFM	67.9	73.5	70.7
fbank + kurt.+SFM	79.1	62.3	70.7

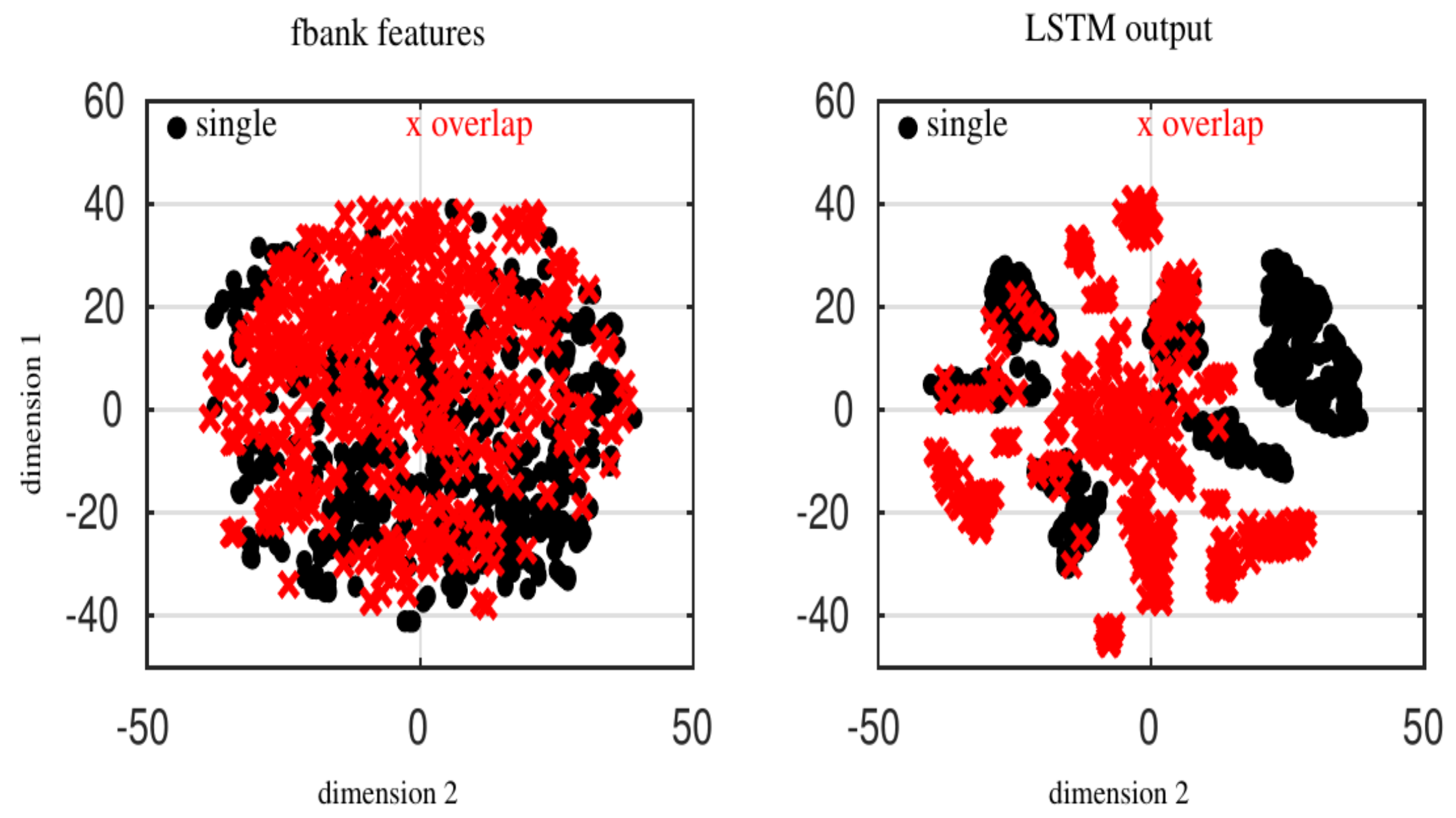

Figure 3: t-SNE scatter plots of input fbank features with context (11×64) and the LSTM 1st layer activation, for single speaker and overlap frames.

Table 4: Detection accuracy %

Model	AMI Dataset (force aligned) with fbank features		
	Single	Overlap	Avg.
DNN[3 layers]	63.9	78.0	70.9
CNN2D[3 layers]	73.0	63.8	68.4
lstm[512 cells]	77.0	68.0	**72.5**
blstm[256 cells]	68.9	75.4	72.1
blstm[512 cells]	57.8	79.0	68.4
Conv[1 layer]-lstm[512]	36.3	87.4	61.8
Conv[3 layer]-lstm[512]	39.3	87.2	63.2
lstm[512 cells][without data aug]	66.37	69.23	67.8
lstm + Viterbi decode	**87.9**	**71.0**	**79.4**

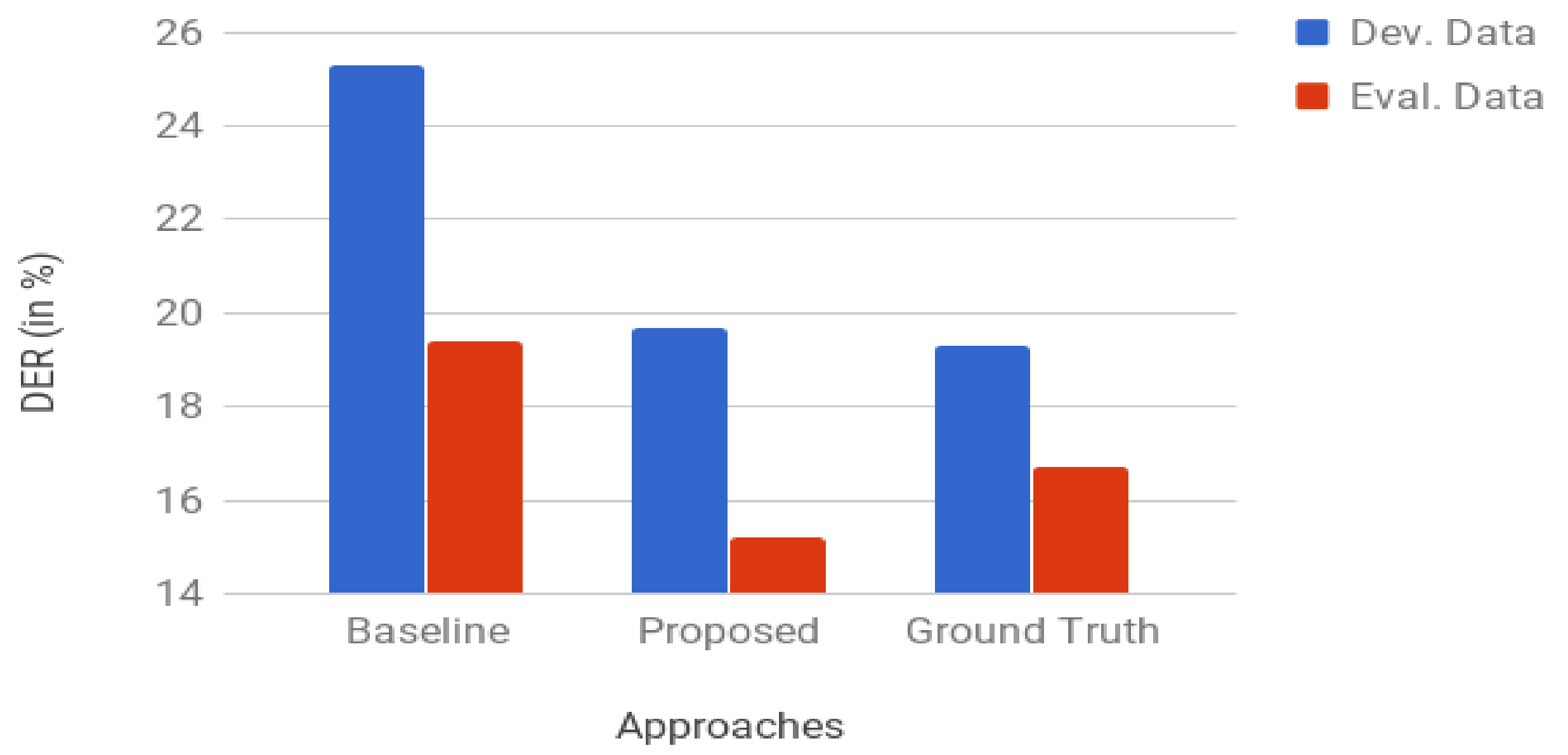

Figure 4: Diarization error rate (DER) % on AMI meetings obtained for different approaches to handle overlap segments.

Acknowledgement

We thank the support from the Jelinek Summer Workshop on Speech and Language Technology (JSALT) 2017 held at CMU, supported by JHU, and the funding from the Defence Research and Development Organisation (DRDO) under the DST0689 project and the Prathiksha Grant.

References

[1] N. Shokouhi et al. "Robust overlapped speech detection and its application in word-count estimation for Prof-Life-Log data". In: *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* Apr. 2015, pp. 4724–4728.

Presented at the 43rd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 15-20 April 2018, Calgary, Alberta, Canada