Discriminative Neural Clustering for Speaker Diarisation

Qiujia Li*, Florian Kreyssig*, Chao Zhang, Phil Woodland

SLT 2021

Cambridge University Engineering Department, Cambridge, UK

Highlights of DNC (I)

- This paper proposes Discriminative Neural Clustering (DNC) as a *supervised* clustering method for diarisation.
- DNC formulates clustering as a sequence-to-sequence task:
  - input: a sequence of segment-level speaker embeddings ($x_{1:N}$)
  - output: a sequence of cluster labels ($y_{1:N}$) mapped from speaker IDs ($z_{1:N}$) based on the order of occurrence, *e.g.*

identity sequence $z_{1:N}$	cluster label sequence $y_{1:N}$
E A C A E E C	1 2 3 2 1 1 3
A C A B B C D B D	1 2 1 3 3 2 4 3 4

- DNC uses attention-based seq2seq models, *e.g.* Transformers.

Highlights of DNC (II)

DNC can su er from data sparsity since a training sample is normally an entire meeting / conversation.

Three data augmentation schemes are proposed:

- sub-sequence randomisation
- input vector randomisation
- Diaconis augmentation

Highlights of DNC (III)

AMI is a very challenging dataset for supervised clustering:

- only 147 meetings for training with 155 di erent speakers
- each meeting has 3-4 spontaneous speakers
- meetings can have up to 1682 segments
- average duration of meetings is 38 min

Together with data augmentation, a curriculum learning approach of training with increasingly long sub-meetings is shown to be e ective.

DNC achieves a reduction in speaker error rate (SER) of 29.4% relative to spectral clustering.

Code is available at https://github.com/FlorianKrey/DNC

Introduction

Clustering: grouping data samples based on similarity.

Speaker diarisation: who spoke when

- segmentation, split audio into speaker-homogeneous segments
- speaker embedding extraction
- clustering of speaker embeddings

Discriminative Neural Clustering (DNC) is applied to the last step of diarisation.

Motivation

Commonly used clustering algorithms:

- e.g. agglomerative / K-means / spectral clustering
- mostly unsupervised & model-free
- leverage pre-de ned distance measures

Clustering is ambiguous & challenging:

- improve speaker embeddings to t clustering algorithms better
- desirable to have a parametric model to learn from examples

Discriminative Neural Clustering (DNC)

Clustering as a sequence-to-sequence problem:

- input: a sequence of speaker embeddings $X = [x_1, \ldots, x_N]^T$
- absolute identity: each x_i has an underlying identity z_i
- output: each x_i is assigned to a cluster label $y_i \in 1, \ldots, N$
  - outputs are relative identities rather than absolute identities
  - labels by the order of occurrence, no permutation required

For example:

identity sequence $z_{1:N}$	cluster label sequence $y_{1:N}$
E A C A E E C	1 2 3 2 1 1 3
A C A B B C D B D	1 2 1 3 3 2 4 3 4

Encoder-decoder Models for Clustering

Encoder-decoder is a discriminative model:

$$p(y_{1:N}|X) = \prod_{i=1}^{N} p(y_i|y_{0:i-1}; X)$$

$$H = \text{Encode}(X)$$

$$y_i = \text{Decode}(y_{0:i-1}; H)$$

Transformer:

Data Augmentation for DNC

Data augmentation:

address data sparsity & avoid trivial solutions

generate as much data as possible vs generate realistic data

Techniques:

(a) random sub-sequence:

using many sub-sequences $(X_{s:e}, y_{s:e})$ of the full sequence

(b) randomisation of input vectors:

Data Augmentation for DNC

(c) Diaconis augmentation (Diac-Aug):

input vectors x_i are L_2-normalised, forming clusters on a hypersphere
Diac-Aug rotates the entire input sequence X to a different region of the hypersphere
a random rotation matrix R is sampled and $X' = XR$

Experimental Setup

AMI meeting corpus, MDM with beamforming.

	#meetings	avg. duration	#speakers
train	147	37.9 min	155
dev	18	32.3 min	21
eval	16	34.0 min	16

32-D Speaker embeddings, trained using angular Softmax

Assume embeddings uniformly distributed on hypersphere

Manual segmentations are used for evaluation

Evaluated using speaker error rate (SER)

Experimental Results (Data Augmentation)

Sub-meeting randomisation is used throughout

Per original meeting 5000 sub-meetings were generated and augmented

randomisation	w/o Diaconis	w/ Diaconis
none	20.19	15.25
global	14.47	19.80
meeting	23.03	13.57

Table 1: SERs for di erent data augmentation techniques for sub-meetings of length 50 segments.

Experimental Results (Curriculum Learning)

Increasing sequence length gradually (N=50 to full length)

Full length meetings use variable N of between 50% and 100%

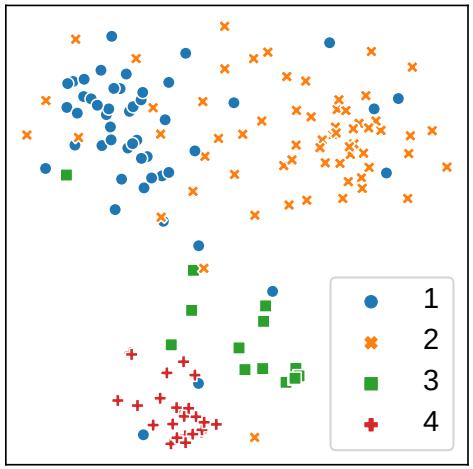
Using meeting-level randomisation & Diaconis augmentation

Finetuned models are initialised from corresponding models using full augmentation; netuned without augmentation.

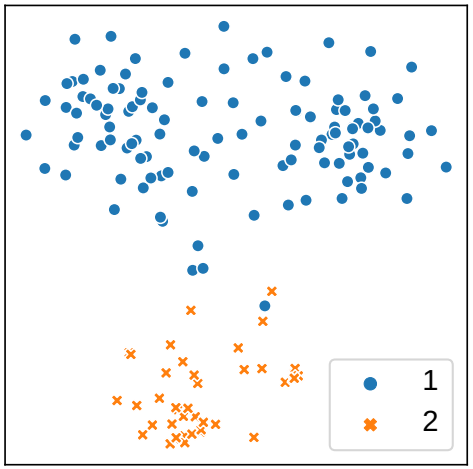
(sub-)meeting length		DNC		SC
#segments	duration	data aug.	netune	
50	2.8 mins	13.57	13.90	15.89
200	9.7 mins	16.92	16.75	22.38
500	20.9 mins	17.73	18.39	23.56
all	34.0 mins	20.65	16.92	23.95

Table 2: Comparison of DNC vs. spectral clustering (SC).

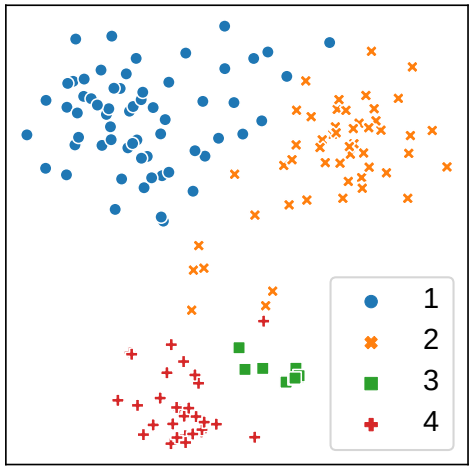
Analysis using t-SNE projection



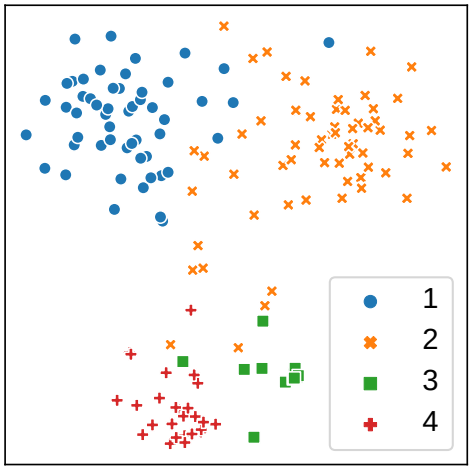
(a) ground truth



(b) spectral clustering (SC)



(c) SC with known #speakers



(d) DNC

Conclusions

- Proposed discriminative neural clustering (DNC), a novel supervised clustering approach using encoder-decoder models.
- Data augmentation & curriculum learning are important.
- DNC performs much better than spectral clustering on a challenging speaker diarisation task.
- DNC can be further extended to a truly end-to-end differentiable diarisation pipeline.