SPEAKER DIARIZATION WITH LSTM

Quan Wang, Carlton Downey, Li Wan, Philip A. Mansfield, Ignacio Lopez Moreno
Google Inc., Carnegie Mellon University
{quanw,liwan,memes,elnota}@google.com

Overview

- We use an LSTM-based speaker verification model [1] for speaker diarization.
- Model is trained on anonymized voice searches, and evaluated on **out-of-domain** data (CALLHOME & NIST RT-03 *etc.*).
- With a modified version of spectral clustering, we achieve state-of-the-art Diarization Error Rate (DER).

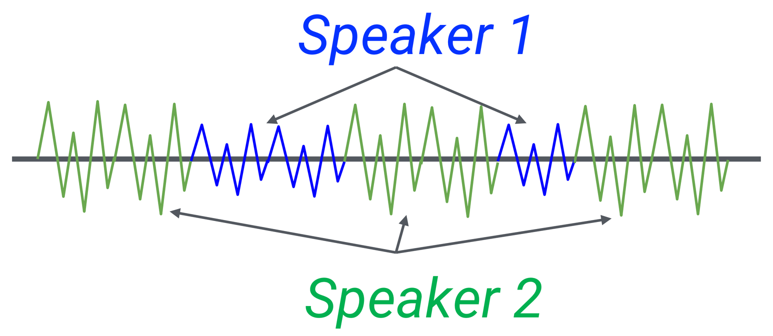


Fig. Speaker diarization solves the problem of "who spoke when".

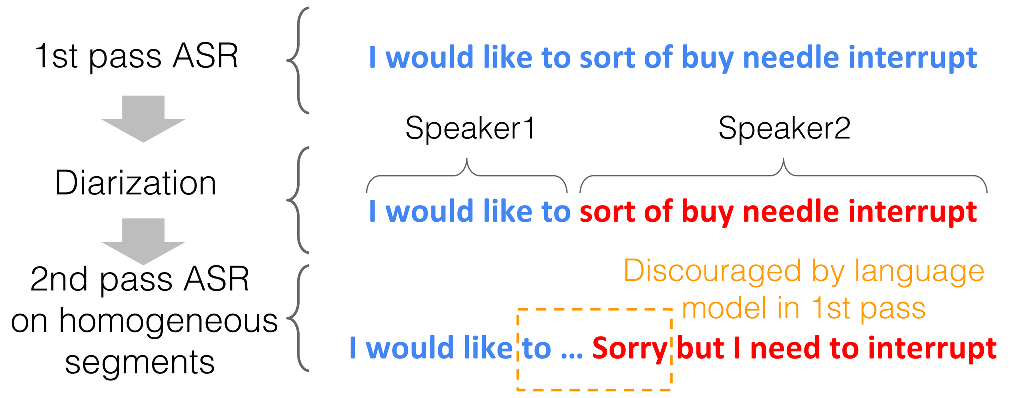


Fig. Example application: Improve ASR with diarization results.

D-Vector Embedding: TE2E to GE2E

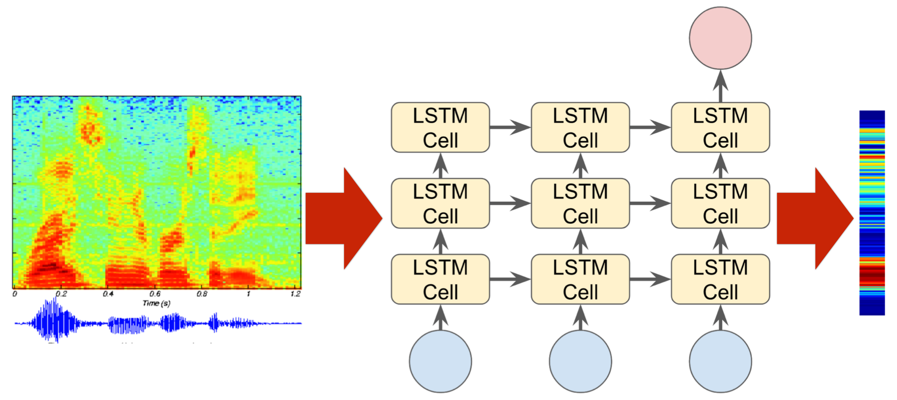


Fig. We use multi-layer LSTM network as audio feature extractor.

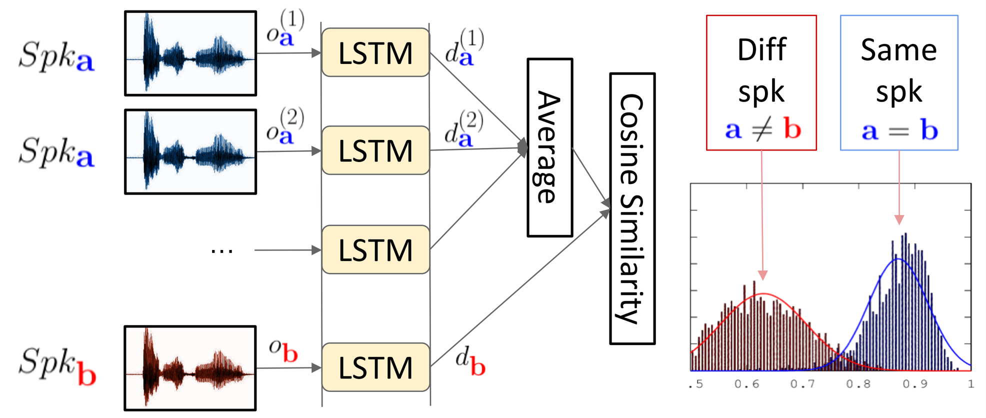


Fig. Tuple E2E loss: Speaker verification as a binary classification problem.

- Tuple E2E [2]: Simulate enroll-verify runtime logic during training. However, tuples are randomly selected, and most tuples are easy – inefficient.
- Generalized E2E loss [1]: For each speaker, focus on its most offensive imposter in the batch.

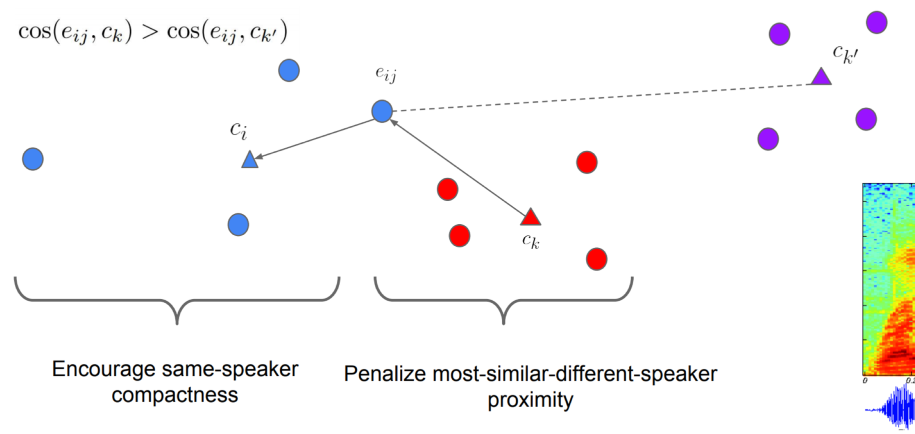


Fig. Generalized E2E: For embedding e_{ij}, we want it to be close to true speaker's center c_i, and distant from the closest false speaker's center c_k.

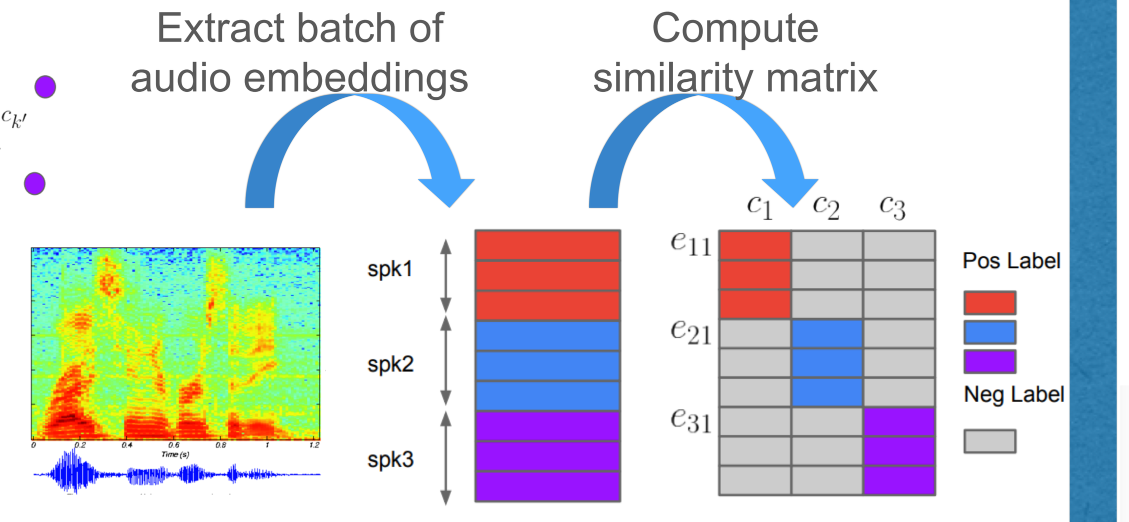
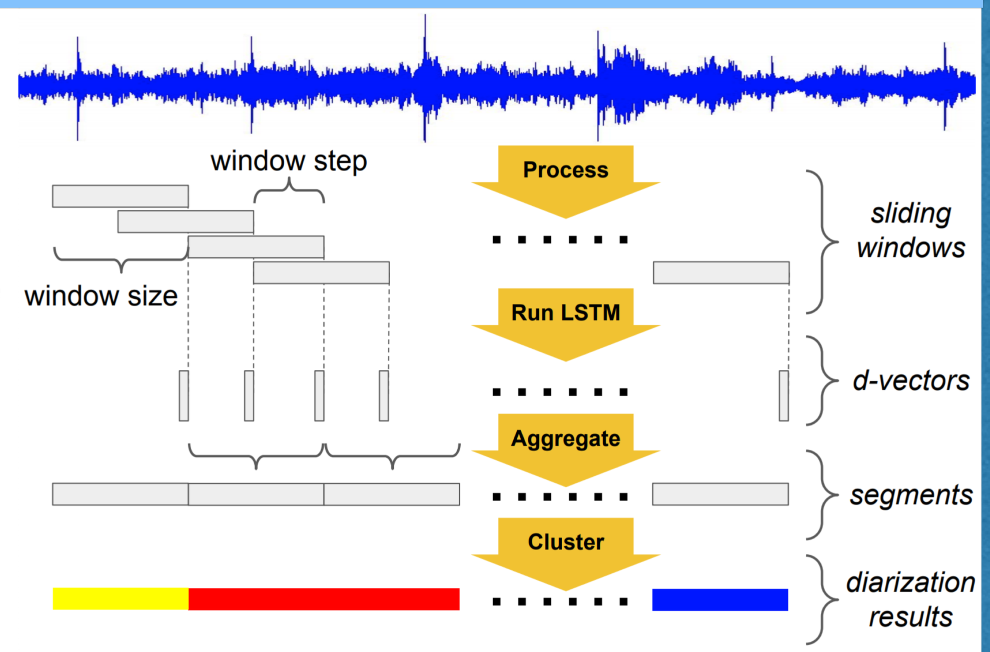


Fig. For each training batch, we build a matrix for utterance-to-speaker similarities, which greatly accelerates the loss computation.

Sliding Window Inference

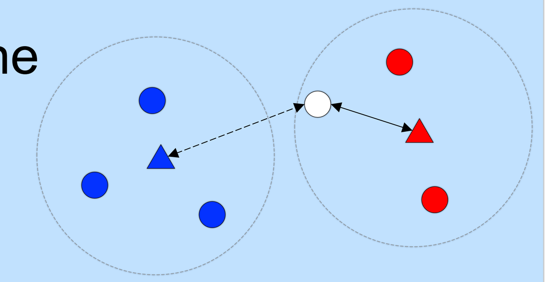
- **Window**: Overlapping, fixed-length (240ms), LSTM runs on it.
- **Segment**: Non-overlapping, longer ($\leq$ 400ms), we average window-wise d-vectors on it.
- Then we cluster segment-wise d-vectors to generate final diarization results.



Clustering Algorithms

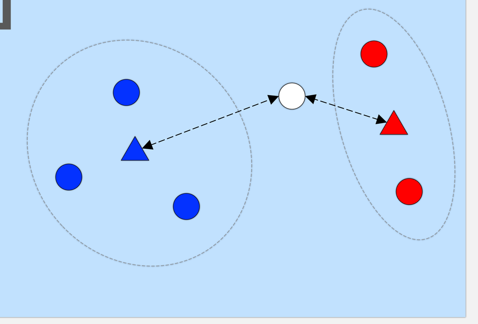
Naïve online

Thresholding the cosine similarities to centroids.



Links online [3]

Anisotropic, probabilistic, and generative cluster modeling.



K-Means offline

- K-Means++ for cluster initialization.
- Find k using Mean Squared Cosine Distances (MSCD):

$$\tilde{k} = \arg\max_{k\geq 1} \mathrm{MSCD}'(k)$$

Spectral offline (Winner!)

- Eigen-decompose affinity matrix.
- Run K-Means on dimensionality-reduced embeddings.
- Find k using the max eigen-gap criterion.

Fig. We experimented with four clustering algorithms. Two of them are online (Naïve and Links), and the other two are offline (K-Means and Spectral).

Affinity matrix refinement: The key to the success of ***spectral clustering***.

- Gaussian blur: Smooth the data, and reduce the effect of outliers.
- Row-wise thresholding: Zero-out affinities between different speakers.
- Symmetrization: Restore matrix symmetry.
- Diffusion: Sharpen affinity section boundaries of distinct speakers.
- Row-wise max normalization: Avoid undesirable scale effects.

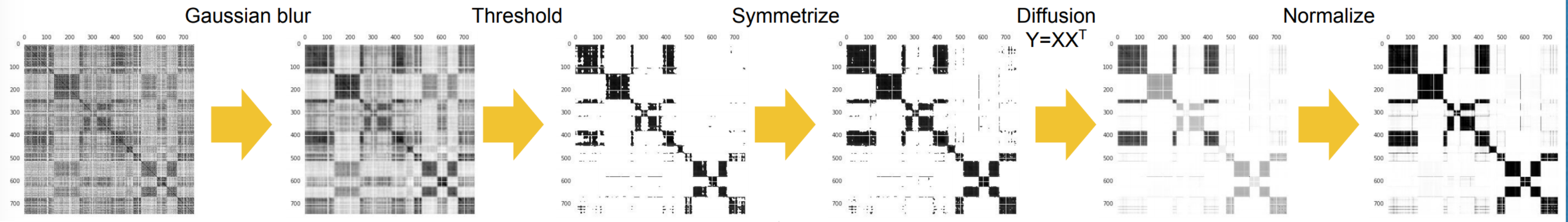


Fig. We apply a sequence of refinement operations on the affinity matrix.

Experiment Results

Training Set
- Anonymized voice searches
- 36M utterances from 18K speakers

→ Train LSTM network

Dev Set
- Separated from Eval Set

→ Tune VAD, window size/step, refinement parameters, *etc.*

Eval Set
- Standard public datasets: NIST RT-03, NIST SRE 2000, *etc.*

→ Report DER

Fig. Our network is completely trained on out-of-domain data: voice search vs. telephone speech.

Embedding	Clustering	CALLHOME American English Eval			NIST RT-03 English CTS Eval		
		Confusion	FA	Miss	Confusion	FA	Miss
i-vector	Naive	26.41	2.40	3.55	35.35	4.66	2.62
	Links	25.40			33.56		
	K-Means	22.86			24.38		
	Spectral	14.59			13.84		
d-vector	Naive	12.41	1.94	4.51	18.76	4.09	4.45
	Links	11.02			18.56		
	K-Means	7.29			7.80		
	Spectral	6.03			3.76		

Table. DER (%) on English-only datasets for different embeddings and clustering algorithms.

Method	Confusion	FA	Miss
Our model	12.0	2.2	4.6
Castaldo [4]	13.7	–	–
Shum [5]	14.5	–	–
Senoussaoui [6]	12.1	–	–
Sell [7] (+VB)	13.7 (11.5)	–	–
Romero [8] (+VB)	12.8 (9.9)	–	–

Method	Confusion	FA	Miss
Our model	5.97	2.51	4.06
Zajíc [9]	7.84	–	–

Table (Left). DER (%) on NIST SRE 2000 CALLHOME. VB for Variational Bayesian resegmentation.

Table (Up). DER (%) on CALLHOME American English 2-speaker subset (CH-109).

References

[1] Li Wan, *et al.*, "Generalized end-to-end loss for speaker verification," *arXiv:1710.10467*, 2017.
[2] Georg Heigold, *et al.*, "End-to-end text-dependent speaker verification," *ICASSP 2016*.
[3] Philip Andrew Mansfield, *et al.*, "Links: A high dimensional online clustering method," *arXiv:1801.10123*, 2018.
[4] Fabio Castaldo, *et al.*, "Stream-based speaker segmentation using speaker factors and eigenvoices," *ICASSP 2008*.
[5] Stephen H Shum, *et al.*, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *TASLP 2013*.
[6] Mohammed Senoussaoui, *et al.*, "A study of the cosine distance-based mean shift for telephone speech diarization," *TASLP 2014*.
[7] Gregory Sell, *et al.*, "Diarization resegmentation in the factor analysis subspace," *ICASSP 2015*.
[8] Daniel Garcia-Romero, *et al.*, "Speaker diarization using deep neural network embeddings," ICASSP 2017.
[9] Zbyněk Zajíc, *et al.*, "Speaker diarization using convolutional neural network for statistics accumulation refinement," *INTERSPEECH 2017*.

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