# End to End Speaker/Speech Diarization with Self-Attention for Conversations

Myat Aye Aye Aung, Win Pa Pa

University of Computer Studies, Yangon

# Outlines

- Background
- Targets
- Proposed Method
- The impact of the proposed system
- Outcomes
- Conclusion

## Background

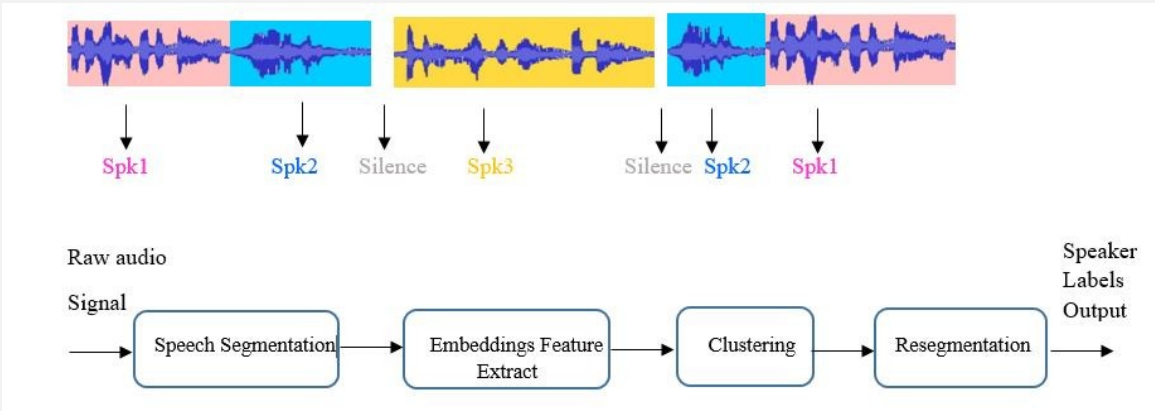



**Figure: Main components of speaker diarization**

- Speaker Diarization is a task to identify “who spoke when”.
- This can be useful for transcribing meetings, classroom speech, or medical interactions.
- It helps AI and humans understand who is saying what throughout the conversation.
- The main problem would be solved by speaker diarization is to determine the number of people involved in a conversation of either in the meeting, broadcast news, or telephone.
- The speaker diarization serves the problem solving for indexing and retrieval-based systems.
- It is to support low resource languages that is an important pre-processing step for many speech applications.

## Targets:

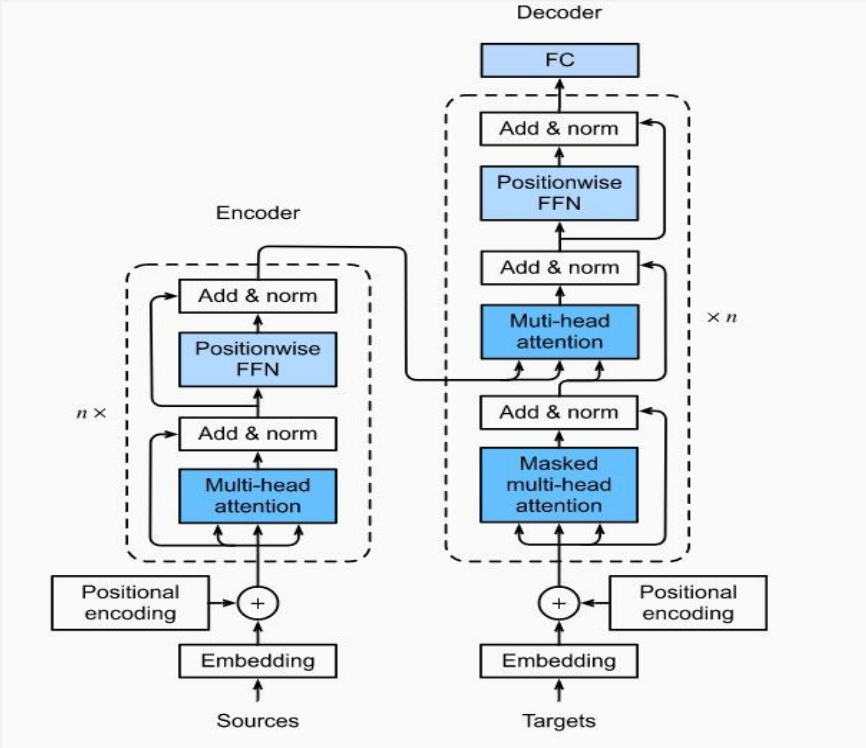
- It has been mainly developed based on the clustering of speaker embeddings.
- Clustering-based approach has problems such as not optimized to minimize diarization errors directly, and cannot handle speaker overlaps correctly.
- End-to-End Neural Diarization (EEND), in which a directly outputs speaker diarization results given a multi-talker recording, was recently proposed.
- In this research, enhance EEND by using self-attention blocks instead of BLSTM blocks.
- In contrast to BLSTM, which is conditioned only on its previous and next hidden states, self-attention is directly conditioned on all the other frames, making it much suitable for dealing with the speaker diarization problem.
- To achieve good performance and proposed method to be significantly better than the conventional BLSTM-based method.
- Speaker Diarization is useful because it makes to increase transcript readability and better understand what a conversation is about.

# Proposed Method: Self-Attention

- Typical speaker diarization systems are based on the clustering of speaker embeddings. For instance, i-vectors, d-vectors, and x-vectors are commonly used in speaker diarization tasks.
- These embeddings of short segments are partitioned into speaker clusters by using clustering algorithms, such as Gaussian mixture models, agglomerative hierarchical clustering, mean shift clustering, k-means clustering, etc.
- This clustering methods have shown themselves to be as effective on various datasets.
- However, such clustering-based methods have a number of problems.
  - cannot be optimized to minimize diarization errors directly
  - have trouble handling speaker overlaps, and
  - have trouble adapting their speaker embedding models to real audio recordings with speaker overlaps

# Proposed Method: Self-Attention (cont'd)

- These problems hinder the speaker diarization application from working on real audio recordings that usually contain overlapping segments.
- To solve these problems, this research has proposed Self-Attentitive End-to-End Neural Diarization (SA-EEND). Different from most of the other methods, the proposed method does not rely on clustering.
- Instead, a self-attention-based neural network directly outputs the joint speech activities of all speakers for each time frame, given an input of a multi-speaker audio recording.
- This method can naturally handle speaker overlaps during the training and inference time by exploiting a multi-label classification framework.




Source: https://d2l.ai/chapter_attention-mechanisms-and-transformers/transformer.html

# The impact of the proposed system

- Can avoid greedy sequential clustering in case of ultra-long sequence, a self-attention-based neural network directly outputs the joint speech activities of all speakers for each time frame, given an input of a multi-speaker audio recording.
- Can provide minimal diarization errors and achieve good speaker-diarization performance for Language Conversations.
- Can help extract important points or action items from the conversation and identify who said that.
- Can identify how many speakers were on the audio.
- Able to supply any kind of Myanmar language meeting with colleagues that speaker diarization will let an audio recording be turned into meaningful notes right after the meeting.
- Can help automatic speech recognition performance in multi-speaker conversation scenarios in meetings and home environments.
- Able to support that is an important process for a wide variety of Myanmar Language applications such as information retrieval from broadcast news, meetings, and telephone conversations, speaker indexing and retrieval, etc.

# Outcomes

- There are several benefits of making good speaker diarization systems such as
  - ✓ Improving existing speech recognition systems,
  - ✓ Making transcripts more meaningful and searchable, or
  - ✓ Assisting hearing impaired people with identifying different speakers on conference calls.
- Speaker Diarization has certain applications in many important scenarios, such as
  - ✓ Understanding medical conversations,
  - ✓ Multimedia information retrieval,
  - ✓ Video Captioning,
  - ✓ Speaker indexing and retrieval,
  - ✓ Speech recognition with speaker identification,
  - ✓ Diarizing meeting and lectures.

# Outcomes

- This research of data sets are real time speech such as interview, panel discussion and talk show data to support existing speech recognition system especially spontaneous speech recognition in low resources.
- The output of this system and technologies to benefit for any language conversations and will collaborate with new partners and colleagues.

# Conclusion

- an advanced topic in speech processing.
- It solves the problem “who spoke when”, or “who spoke what”.
- It is highly relevant with many other techniques, such as voice activity detection, speaker recognition, automatic speech recognition, speech separation, statistics, and deep learning.
- There are also challenges arising in the speaker diarization area such as number of speakers, speech activity detection, amount of overlap.
- End-to-End Neural Diarization is a neural network for speaker diarization in which a neural network directly outputs speaker diarization results given a multi-speaker recording.
- The end-to-end method can explicitly handle overlaps during training and inference.
- It feed multi-speaker recordings with corresponding speaker segment labels, the model can be adapted to real conversation.
- It could be believed that if the end-to-end method is used in this research, the experimental result will be good and can supply as a pre-processing step for various speech processing applications.

# Thank You!