Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M18-209P^
Published: 2/26/2020

Inventors

Huan Song
Visar Berisha
Andreas Spanias
Megan Willi
Jayaraman Thiagarajan

Contact

Shen Yan
shen.yan@skysonginnovations.com

Triplet Network with Attention for Speaker Diarization

Background

In automatic speech processing systems, speaker diarization is a crucial front-end component for separating speech segments by speaker without a priori knowledge about speaker identities. The first phase of many state-of-the-art diarization techniques involves the conversion of original speech data into representative i-vectors, which are then scored in fully connected networks for metric learning. Despite adequate result quality, these approaches have lagged in their adaptation to current trends in deep neural networks (DNN), notably in terms of joint representation and task-based learning. Redesigning the diarization process to fully leverage DNN architecture stands to accelerate development of robust and reliable solutions.

Invention Description

Researchers at Arizona State University have developed a new process for speaker diarization. Dispensing entirely of i-vector formation, this process employs a deep attention network paired with a triplet ranking loss algorithm. Bypassing the resource-intensive i-vector extraction step significantly reduces training time and required training data.

Embeddings from the Mel-frequency cepstral coefficients (MFCCs) and speech similarity metrics are jointly learned directly from raw speech segments. A multi-head, self-attention mechanism forms the basis for the attention model, while a triplet ranking loss is trained with input sets that each contain three samples: a positive, a negative, and an anchor. Anchor-to-positive distances and anchor-to-negative distances are computed and evaluated against a local margin value. Unlike global margin values used in some contrastive loss techniques, the local margin value affords this model more flexibility and expressive power.

Training experiments using the CALLHOME corpus resulted in a diarization error rate (DER) of 12.7% with the new method, while DERs for conventional i-vector techniques ranged from 13.4% to 18.7%.

Potential Applications

- Speech processing
- Deep neural networks
- Machine learning

Benefits and Advantages

- Resource-saving – Avoiding i-vector construction reduces required training time and data
- Flexible – Triplet loss formulation features a local margin value for added expressive power
- Effective – Outperforms i-vector techniques in experiments
- Integrative – Adds deep neural network features to speech diarization

Related Publication (PDF)

Hompage of Professor Andreas Spanias