Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

# A Hybrid Approach to Combining Role Diarization with ASR for Professional Conversations

*Bongjun Kim, Arindam Ghosh, Mark Fuhs, Anurag Chowdhury, Deblin Bagchi, Monika Woszczyna*

Solventum Health Information Systems, USA

{bkim11, aghosh4, mark.fuhs, achowdhury2, dbagchi, mwoszczyna}@solventum.com

## Abstract

In professional settings, conversations often involve persons with defined roles (doctor, patient, lawyer, client, etc.), and the intelligibility of a conversational transcript may be improved by annotating conversational turns with the role of the speaker, e.g. "Doctor: How are you feeling? Patient: I sprained my ankle." We propose a novel hybrid architecture that combines an ASR model augmented to label the speaker's role at each speaker change point with a d-vector-based diarization system. This system outperforms modular and fully integrated baselines by 12% and 28%, respectively. We also show that, when an ASR transducer model is trained to predict role or speaker-change tokens as part of the transcript, these token timings can improve diarization more than the adjacent word token timings can, despite there being no explicit training signal conveying precise speaker change points.

**Index Terms**: role diarization, speaker diarization, speech recognition

## 1. Introduction

Speaker diarization is the process of inferring which speaker spoke when in a multi-party conversation. Often, the problem is approached from the standpoint that the number of speakers and their identities are not known, so the produced speaker labels are anonymous: speaker-1, speaker-2, etc. Other work has explored tying these anonymous labels to known speaker identities using external speaker profiles. However, in many types of professional conversations, it is valuable to identify the *roles* of the speakers in the conversation – doctor vs. patient, professional services provider (lawyer, accountant, etc.) vs. client(s), call-center agent vs. caller, etc. – even in the absence of background information about the identities of the conversation's participants.

For many utterances, the role of the speaker may be inferred from the transcript text alone [1, 2]. For example, in the medical domain, utterances such as "how long has your back been hurting" strongly indicate that the speaker is a medical provider, not a patient. Recent work [2] has proposed using LLMs to diarize the transcribed text with role labels. However, text is not always a reliable indicator, particularly during rapid-fire exchanges or portions of professional conversations where the intent is to build rapport. Rapport building during medical encounters may include discussion of hobbies, sports and other general topics that may nonetheless play a useful role in medical treatment [3].

Several studies have explored the problem by leveraging both acoustic and linguistic information [4, 5, 6, 7, 8] for role diarization. In [6], modular approaches are employed whereby utterances are first textually classified into roles using role-specific language models, then reclassified based on acoustic x-vector centroids for each role. Subsequent work by [7] expands on this strategy, using a BERT-based text classifier for role classification, followed by a role-constrained clustering of x-vectors.

By contrast, a fully integrated approach was proposed in [9], where an RNN transducer, a traditionally ASR-oriented architecture, was trained to infer both the spoken text and role identity by augmenting the training text with special role tokens for doctor and patient; other speakers were grouped into one of those two roles. No information across speech segments was used. In [8], a similar fully integrated model trained for air-traffic control conversations was compared with separate models for ASR and role diarization, where the role diarization was based either on only acoustic or only linguistic information.

Though speaker and role diarization are fundamentally different problems, some of the recent work in speaker-attributed ASR (SA-ASR), where systems jointly infer what was spoken and who spoke it, is relevant as it inspires potential solutions to the problem of role-attributed ASR. [9] was extended to a more general speaker diarization scenario by [10], who embedded a generic turn token between speaker changes in the transcript. After resegmenting the audio using the turn token timestamps, a traditional speaker encoder model was then used to generate d-vectors which were clustered using a speaker-turn-constrained version of spectral clustering.

Recent SA-ASR efforts [11, 12, 13, 14, 15] have also explored how to fuse ASR and speaker attribution into a single neural architecture. In [15], an auxiliary speaker diarization network was combined with a pretrained and frozen ASR transducer model to remove any ASR performance degradation due to optimizing multiple training losses. For coherent speaker labels, the whole conversation is processed at once, but this leads to low accuracy on conversations longer than a few minutes, the norm in professional settings.

The contributions of this paper are two-fold. First, we propose a new architecture for ASR + role diarization and show that it outperforms both modular and fully-integrated baseline systems on a large and challenging real-world dataset from the domain of far-field doctor-patient conversations. Second, we analyze the relative strengths of these three systems and several additional system variants to better understand the factors that contribute to accurate role diarization.

## 2. Architectures for Role Diarization

First, we describe modular and fully-integrated baseline architectures for performing ASR and role diarization. We then propose a hybrid architecture and several variants that provide comparative insights. For all systems, a TDNN-LSTM-based voice



10.21437/Interspeech.2025-895

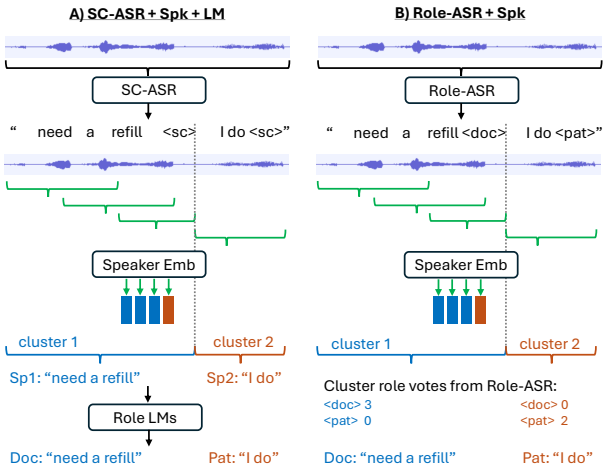


Figure 1: *Step-by-step inference process, illustrated for a single VAD segment, in the (a) SC-ASR + Spk + LM and (b) Role-ASR + Spk architectures.*

activity detector is first applied to segment an entire recording into silence-separated segments, typically less than 18 sec long.

### 2.1. Modular Baseline: SC-ASR+Spk+LM

For this system, speech recognition, speaker diarization and role labeling are all performed by separate systems. Figure 1a shows the following stages in the inference process, illustrated for a single VAD segment:

**ASR:** Speech recognition is performed using an ASR transducer model trained to output the spoken text of the conversation as well as a speaker-change token ($\langle$sc$\rangle$) at the end of each speaker's turn, similar to [10]. ASR token timestamps are derived from the encoder frame associated with each transducer token. The output is a transcript containing segments $S_i^a = (t_i^{start}, t_i^{end}, textsc_i)$ with VAD time boundaries and $textsc_i = [(tok_0, t_0^{tok}), (tok_1, t_1^{tok}), ...]$, containing the spoken words and speaker-change tokens within the segment.

**Resegmentation:** The segments are divided at the speaker-change token times, creating new segments $S_j^r = (t_j^{start}, t_j^{end}, text_j)$ that contain the spoken word tokens for a single speaker turn.

**Diarization:** The resegmentation is used to create speaker embeddings (d-vectors) generated by an ECAPA-TDNN model. A d-vector is generated for each 1.5 sec window, shifted across each new segment in 250ms increments. All d-vectors from the recording are clustered, yielding a speaker diarization of the conversation. The result is a new transcript comprising segments $S_k^d = (t_k^{start}, t_k^{end}, s_k, text_k)$, where $s_k$ is an absolute speaker index. There may be more segments $S_k^d$ than input segments $S_j^r$ if d-vectors from the same input segment end up in different speaker clusters, which is permitted since speaker change tokens have higher precision than recall.

The effect of the resegmentation on the diarization stage is illustrated in Figure 1a. The speaker-change token divides (vertical dotted black line) the VAD segment, creating two smaller segments within which spans of audio are selected (green brackets) to be used for d-vector generation. By dividing the VAD segments at speaker changes, each d-vector is less likely to span across multiple speakers. The improved speaker "purity" of the resulting d-vectors leads to higher cluster quality and ultimately lower diarization error rates.

**Roles:** Three role-specific LMs (Doctor, Patient and Other), trained to maximize the likelihood of word sequences for a particular role, are applied to the text of each speaker segment, and the lowest-perplexity role is assigned to the words in that segment. Then, role votes are tallied to assign a role to each speaker. The final output is a role diarized transcript with segments $S_k^{rd} = (t_k^{start}, t_k^{end}, r_k, text_k)$, where $r_k$ indicates a role label: Doctor, Patient, OtherPerson1, OtherPerson2, etc.

For this *and all other architectures that use d-vector-based diarization*, roles are assigned to speakers by tallying role votes across all of the words in all of that speaker's segments. The speaker with the most votes for Doctor is chosen first. Next, the speaker with the most votes for Patient is chosen. However, if the percentage of votes for Other exceeds a threshold, that speaker is not labeled as the Patient. This threshold, tuned per system on a validation set, is typically only 25-35% in our data domain, since caregivers in the conversation may speak on behalf of the patient using highly patient-like language. Each additional speaker in the conversation not assigned a known role is given a unique label, e.g. "OtherPerson1". Section 3.3 describes how this is handled by the scoring metrics.

### 2.2. Fully Integrated Baseline: Role-ASR

Similar to [9] and in contrast with the modular baseline system above, this is a fully integrated architecture in which an ASR transducer model is trained to output both spoken text and special role tokens ($\langle$doc$\rangle$, $\langle$pat$\rangle$, $\langle$other$\rangle$) at the end of each speaker's turn:

**ASR:** The Role-ASR system is used to produce a transcript containing segments $\{S_1^a ... S_n^a\}$, where each segment $S_i^a = (t_i^{start}, t_i^{end}, textrl_i)$ comprises the VAD time boundaries and $textrl_i = [(tok_0, t_0^{tok}), (tok_1, t_1^{tok}), ...]$ contains the spoken words and role tokens within the segment.

**Resegmentation:** The segments are divided at role token times and labeled by the role tokens, creating a role diarized transcript containing segments $S_j^{rd} = (t_j^{start}, t_j^{end}, r_j, text_j)$, where $r_j$ indicates a role label: Doctor, Patient, Other. The system is not able to distinguish among multiple Other speakers which appear in 3% of conversations in our data.

This approach is vastly simpler than the modular approach and allows for end-to-end optimization. However, ASR $x$former-transducer architectures typically process a single VAD segment at a time due to the high resource costs associated with processing longer sequences (quadratic cost for attention; multiplicative cost of audio and transcript lengths for transducer loss).

### 2.3. Proposed: Role-ASR+Spk

We propose a hybrid architecture, illustrated in Figure 1b, in which a Role-ASR system is coupled to a speaker diarization system in a way that combines the benefits of integrated speaker change detection and role classification within the Role-ASR system with the global data view of d-vector-based speaker diarization:

**ASR & Resegmentation:** As with the Role-ASR baseline.

**Diarization:** The role labels are set aside at this stage and d-vector-based diarization is performed on the resegmented transcript segments, resulting in a new transcript segmentation with both speaker and role information: $S_k^d = (t_k^{start}, t_k^{end}, r_k, s_k, text_k)$.

**Roles:** For each segment, the role label $r_k$ from the Role-ASR system is used to vote on the role of the corre-

sponding speaker, $s_k$, weighted by the number of words in $text_k$. Then, role votes are tallied to assign a role to each speaker. The final role diarized transcript has segments $S_k^{rd} = (t_k^{start}, t_k^{end}, r_k, text_k)$, where $r_k$ indicates a role label: Doctor, Patient, OtherPerson1, OtherPerson2, etc.

### 2.4. Role-ASR+Spk Alternative Formulations

In order to better understand the interaction between the Role-ASR role tokens and the diarization system, we also consider three alternatives:

**Without resegmentation**: In this alternative, the VAD segmentation is provided to the diarization system without dividing the VAD segments based on role token timings, i.e. without the benefit of speaker change detection. But role classification of speakers continues to be based on the Role-ASR system.

**With word-based speaker changes**: Instead of using the timings of the role tokens to indicate time points of speaker change, we explored several different change-point estimates based on the times of the token of the last word before a role token and the first word token thereafter. Our best performing strategy involved setting the change point to be at the midpoint between the last token before the role token and the first token after the role token.

**Role-ASR+Spk+LM**: This alternative uses the Role-ASR system with an output filter to map each role token to a speaker-change token, thus leveraging the speaker-change information in the role tokens while still using LMs for role classification.

## 3. Experiments

### 3.1. Data

All models were trained and tested on anonymized hand-transcribed recordings of doctor-patient conversations (DoPaCo) covering multiple medical specialties. Recordings come from medium-to-far-field microphone locations in medical exam rooms. Speaker identities were transcribed according to their functional role (doctor, patient, caregiver, spouse, other, ...), which we collapse to Doctor, Patient and Other for these experiments. The data were partitioned into training, validation and test sets of 1750h, 26.5h and 35.0h, respectively.

### 3.2. Models

The features for the ASR transducer model were 64-dim log-Mel features computed over a 25ms window with 10ms frame shift. The ASR transducer model encoder was composed of two convolutional subsampling layers that together reduced the frame rate by a factor of 4, followed by 12 e-branchformer layers [16] with dimension 384, 6-headed MHSA and 1536-dim feed-forward projection sublayers. The predictor network was a single-layer 384-dim LSTM network, and the joint network was composed of feed-forward projection networks first to 512-dim, then to the final 5000-token BPE vocabulary, including the blank symbol and role tokens. SpecAugment was applied to the input features during training, and the final models were obtained by averaging weights from the 10 best epochs based on validation loss.

The ECAPA-TDNN speaker embedding model [17, 18] was first trained on VoxCeleb2, then adapted to the DoPaCo data, which we found performed the best among several data combination strategies. For the most part, the ECAPA-TDNN training recipe followed the default recipe given in the SpeechBrain toolkit. However, we replaced the AAM-Softmax loss with the GE2E loss [19] for training the ECAPA-TDNN on DoPaCo data as it provided significant performance improvements. Speaker embeddings are clustered using a simplified variant of VBx clustering [20]. With the improved performance of more modern speaker embedding architectures such as ECAPA-TDNN, PLDA was found not to improve performance. The simplification uses a unit-normal distribution with speaker-specific mean for each HMM state emission distribution. The means are initialized from an agglomerative hierarchical clustering (AHC) of all speaker embeddings in the recording, as described in [20].

The role LMs were 3-layer 128-dim LSTM models trained on the same BPE vocabulary as the ASR model. A single model was first trained on all DoPaCo data, then adapted to either a Doctor-, Patient-, or Other-role-specific subset of the data until perplexity was minimized on the validation set.

### 3.3. Scoring

All scoring statistics are based on multi-speaker alignments computed using *asclite* from SCTK, the NIST Scoring Toolkit. Word error rate (WER) and word diarization error rate (WDER) are computed in the standard way. In the latter case, like diarization error rate (DER), all permutations of assignment of hypothesis speakers to reference speakers are tried, and the highest-scoring assignment is selected.

For role-based word diarization error rate (RWDER), the hypothesis is expected to identify each unique speaker as well as identify which speakers are the doctor and patient. Other speakers are variously labeled "OtherPerson1", "OtherPerson2", etc. RWDER computation is similar to WDER, except that doctor and patient speakers are excluded from the speaker assignment permutation search: the hypo speaker labeled as the doctor must be paired with the reference doctor; as well for the patient. Any additional speakers participate in the permutation search. Thus, RWDER is always higher than WDER.

## 4. Results

### 4.1. Baseline Systems and Role-ASR+Spk

The ASR (WER), speaker diarization (WDER) and role diarization (RWDER) performance of all systems are shown in Table 1, and role confusion matrices are shown in Table 2.

The Role-ASR model (Table 1, system 3), used by itself, exhibits higher WDER and RWDER than the modular baseline SC-ASR+Spk+LM. The Role-ASR model processes each VAD segment separately, and, in this fully integrated approach, role labels are final; there is no global stage to leverage speaker information across VAD segments. In particular, Table 2(b) shows that it more frequently confused Other speakers (typically a caregiver or spouse) with the Patient. From a single segment, there may be few clues to distinguish these roles, even for a human evaluator.

The SC-ASR+Spk+LM system's use of diarization over the whole conversation achieved substantially lower WDER (Table 1, system 1) than the standalone Role-ASR model and ultimately lower RWDER after the LM-based role classification stage. An examination of the outputs of the two systems revealed different patterns of errors. For a speaker whose voice sounded distinct from others in the conversation, the Role-ASR model was more likely to assign some fraction of their utterances to the wrong role. The SC-ASR+Spk+LM system was more likely to err on utterances where the speaker's voice sounded dissimilar from other utterances. This was particularly

Table 1: *Word error rate (WER), word diarization error rate (WDER) and role-based word diarization error rate (RWDER) for the baseline systems (1, 3), the proposed architecture (6), and other variants (2, 4, 5). (SD: Speaker diariazation using d-vector)*

| | | Architecture | ASR Speaker Change | SD | Roles | WER | WDER | RWDER | RWDER - WDER |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 1 | SC-ASR+Spk+LM | SC tokens | Y | LM | 16.6 | 3.8 | 5.4 | +1.6 |
| | 2 | Role-ASR+Spk+LM | role → SC tokens | Y | LM | 16.3 | 3.7 | 5.4 | +1.7 |
| | 3 | Role-ASR | - | N | Role-ASR | 16.3 | 6.1 | 6.6 | +0.5 |
| Ours | 4 | Role-ASR+Spk | not used | Y | Role-ASR | 16.3 | 6.0 | 6.9 | +0.9 |
| | 5 | Role-ASR+Spk | word timings | Y | Role-ASR | 16.3 | 4.6 | 5.5 | +0.9 |
| | 6 | Role-ASR+Spk | role tokens | Y | Role-ASR | 16.3 | **3.7** | **4.7** | +1.0 |

Table 2: *Role confusion matrices of each architecture. Each row represents a reference role, and each column indicates the system's role classification; italics indicates correct hypothesis labels. WrongOther indicates a speaker diarization error among other-role speakers.*

| Reference | Doctor | Patient | Other | WrongOther |
|---|---|---|---|---|
| Doctor | *97.0* | 2.2 | 0.9 | - |
| Patient | 3.5 | *94.5* | 2.0 | - |
| Other | 4.2 | 40.7 | *54.3* | 0.8 |

(a) SC-ASR+Spk+LM

| Reference | Doctor | Patient | Other | WrongOther |
|---|---|---|---|---|
| Doctor | *97.3* | 2.4 | 0.3 | - |
| Patient | 5.5 | *92.5* | 2.0 | - |
| Other | 13.6 | 50.0 | *36.0* | 0.4 |

(b) Role-ASR

| Reference | Doctor | Patient | Other | WrongOther |
|---|---|---|---|---|
| Doctor | *97.0* | 2.1 | 0.9 | - |
| Patient | 3.4 | *94.1* | 2.5 | - |
| Other | 4.2 | 20.7 | *74.2* | 0.9 |

(c) Role-ASR+Spk (role tokens)

impactful when the wrong number of speakers was estimated.

The proposed Role-ASR+Spk system (Table 1, system 6) achieves the lowest RWDER, beating both baseline systems. While the SC-ASR+Spk+LM architecture outperformed the Role-ASR model on RWDER, this can be attributed to the better speaker diarization (lower WDER) upon which the role classification was applied. Seen more directly, the Role-ASR+Spk+LM alternative (system 2) maps role tokens to speaker change tokens and then applies the same speaker diarization and LM-based role classification as in the SC-ASR+Spk+LM architecture. All three architectures – SC-ASR+Spk+LM, Role-ASR+Spk+LM and Role-ASR+Spk – have nearly identical WDER, but Role-ASR+Spk+LM's RWDER regresses to the performance of the SC-ASR+Spk+LM architecture. The power of the Role-ASR model's role classification is indeed stronger than the LM-based approach when aggregated and applied to the speaker diarization of the whole conversation.

### 4.2. Role Tokens and Speaker Diarization

We explored the performance of several alternative systems in order to better understand the contributions of the role tokens to speaker diarization. First, as discussed above, when the Role-ASR model's tokens are treated as speaker change tokens (Table 1, system 2), WDER performance is equivalent to using the SC-ASR model's speaker-change tokens. The resegmentation guided by the role tokens was as beneficial to speaker diarization as the resegmentation guided by speaker-change tokens.

By contrast, the performance of an alternative architecture in which the speaker diarization system is applied to the VAD segments, not the role-token-based resegmentation, is shown in Table 1 as system 4. WDER substantially increases, showing the contribution of temporally precise change-point detection for a speaker diarization system that otherwise uses wide, equally spaced sliding windows of audio to generate speaker embeddings [21, 10].

One subtle aspect of the training of the SC-ASR and Role-ASR systems is that only the normal ASR training losses were used: transducer loss + secondary CTC loss on the encoder. No explicit speaker change timing information is provided as a training signal, so a role or speaker change token produced at any time after the last word of one speaker and before the first word of the next speaker satisfies the loss.

We therefore expected that the precise timing of speaker changes could be estimated just as well using the timing information of the word tokens adjacent to the role tokens. However, despite trying several different strategies, our best performing strategy using word token times to estimate speaker change points fell short. Table 1, system 5, shows the degradation observed when using word timings to estimate speaker change points. WDER degrades below the performance of the systems that use special tokens related to speaker changes (systems 1, 2 and 6). The additional diarization errors then propagated to higher RWDER.

While the SC-ASR- and Role-ASR-based resegmentations provide a substantial benefit to WDER, there is a small cost in WER. An ASR system trained without speaker-change or role tokens achieved a WER of 16.1, or $\sim$1% rel improvement.

## 5. Conclusions

We propose a novel hybrid architecture that combines a Role-ASR system with a diarization system to jointly perform ASR and role diarization, and we show that this approach outperforms the RWDER of the fully integrated baseline by 28% rel and the modular baseline by 12% rel. The hybrid architecture simplifies the modular baseline by not requiring the training of a separate role classification stage. We also show that the SC-ASR and Role-ASR systems are learning and expressing information about speaker change timings in their speaker change or role tokens that improve diarization performance without an explicit training signal to do so. Future work will explore additional architectures to unify the Role-ASR and speaker embedding models, and other architectures inspired by advances in speaker-attributed ASR models to include role attribution.

# 6. References


[1] J. Zuluaga-Gomez, S. S. Sarfjoo, A. Prasad, I. Nigmatulina, P. Motlicek, K. Ondrej, O. Ohneiser, and H. Helmke, "Bertraffic: Bert-based joint speaker role and speaker change detection for air traffic control communications," in *Proc. IEEE SLT*, 2023.

[2] Q. Wang, Y. Huang, G. Zhao, E. Clark, W. Xia, and H. Liao, "DiarizationLM: Speaker Diarization Post-Processing with Large Language Models," *arXiv preprint arXiv:2401.03506*, 2024.

[3] Y. Tian, J. Zhang, H. Cheand, and Y. Sucorresponding, "Weaving rapport: doctors' strategies towards patients' noncompliance," *BMC Health Services Research*, 2023.

[4] F. Valente, D. Vijayasenan, and P. Motlicek, "Speaker diarization of meetings based on speaker role n-gram models," in *Proc. IEEE ICASSP*, 2011.

[5] A. Sapru, S. H. Yella, and H. Bourlard, "Improving speaker diarization using social role information," in *2014 IEEE ICASSP*, 2014.

[6] N. Flemotomos, P. Georgiou, and S. Narayanan, "Linguistically aided speaker diarization using speaker role information," in *The Speaker and Language Recognition Workshop (Odyssey 2020)*. ISCA, 2020. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2020-17

[7] N. Flemotomos and S. Narayanan, "Multimodal Clustering with Role Induced Constraints for Speaker Diarization," in *Proc. Interspeech*, 2022.

[8] A. Blatt, A. Krishnan, and D. Klakow, "Joint vs Sequential Speaker-Role Detection and Automatic Speech Recognition for Air-traffic Control," in *Proc. Interspeech*, 2024.

[9] L. E. Shafey, H. Soltau, and I. Shafran, "Joint Speech Recognition and Speaker Diarization via Sequence Transduction," in *Proc. Interspeech*, 2019.

[10] W. Xia, H. Lu, Q. Wang, A. Tripathi, Y. Huang, I. L. Moreno, and H. Sak, "Turn-to-Diarize: Online Speaker Diarization Constrained by Transformer Transducer Speaker Turn Detection," in *Proc. IEEE ICASSP*, 2022.

[11] L. Lu, N. Kanda, J. Li, and Y. Gong, "Streaming Multi-talker Speech Recognition with Joint Speaker Identification," in *Proc. Interspeech*, 2021.

[12] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng, X. Wang, Y. Gaur, Z. Chen, J. Li, and T. Yoshioka, "Streaming speaker-attributed asr with token-level speaker embeddings," in *Proc. Interspeech*, 2022.

[13] Y. Huang, W. Wang, G. Zhao, H. Liao, W. Xia, and Q. Wang, "Towards word-level end-to-end neural speaker diarization with auxiliary network," 2023. [Online]. Available: https://arxiv.org/abs/2309.08489

[14] D. Raj, M. Wiesner, M. Maciejewski, L. P. García-Perera, D. Povey, and S. Khudanpur, "On Speaker Attribution with SURT," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024.

[15] Y. Huang, W. Wang, G. Zhao, H. Liao, W. Xia, and Q. Wang, "On the Success and Limitations of Auxiliary Network Based Word-Level End-to-End Neural Speaker Diarization," in *Proc. Interspeech*, 2024.

[16] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023.

[17] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020.

[18] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "ECAPA-TDNN embeddings for speaker diarization," in *Proc. Interspeech*, 2021.

[19] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE ICASSP*. IEEE Press, 2018.

[20] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, 2022.

[21] R. Yin, H. Bredin, and C. Barras, "Speaker Change Detection in Broadcast TV Using Bidirectional Long Short-Term Memory Networks," in *Proc. Interspeech 2017*, 2017, pp. 3827–3831.