INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multi-Stream Extension of Variational Bayesian HMM Clustering (MS-VBx) for Combined End-to-End and Vector Clustering-based Diarization

Marc Delcroix[1], *Naohiro Tawara*[1], *Mireia Diez*[2], *Federico Landini*[2], *Anna Silnova*[2], *Atsunori Ogawa*[1], *Tomohiro Nakatani*[1], *Lukas Burget*[2], *Shoko Araki*[1]

[1]NTT Corporation, Japan, [2]Brno University of Technology, Speech@FIT

marc.delcroix@ieee.org

Abstract

Combining end-to-end neural speaker diarization (EEND) with vector clustering (VC), known as EEND-VC, has gained interest for leveraging the strengths of both methods. EEND-VC estimates activities and speaker embeddings for all speakers within an audio chunk and uses VC to associate these activities with speaker identities across different chunks. EEND-VC generates thus multiple streams of embeddings, one for each speaker in a chunk. We can cluster these embeddings using constrained agglomerative hierarchical clustering (cAHC), ensuring embeddings from the same chunk belong to different clusters. This paper introduces an alternative clustering approach, a multi-stream extension of the successful Bayesian HMM clustering of x-vectors (VBx), called MS-VBx. Experiments on three datasets demonstrate that MS-VBx outperforms cAHC in diarization and speaker counting performance.

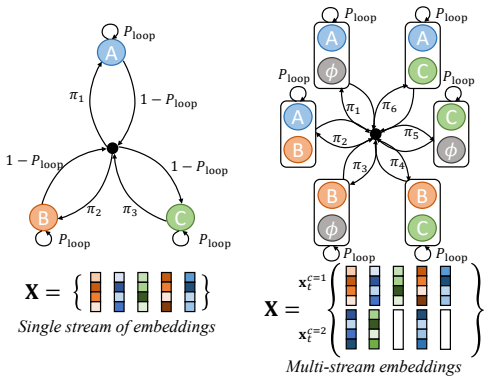
Index Terms: speaker diarization, end-to-end, VBx, clustering

1. Introduction

Diarization consists of determining who speaks when in a multi-talker recording. It plays an essential role in the processing of conversations. There are several approaches to tackle the diarization problem, such as *vector clustering (VC)* [1], *end-to-end diarization (EEND)* [2, 3] and *target speaker voice activity detection (TS-VAD)* [4]. Recently, the combination of the first two, i.e., EEND combined with VC (EEND-VC) has received increased interest since it offers a principled way of getting the best of both frameworks achieving high performance on several tasks [5–7]. The term EEND-VC was introduced in [5], but here it refers to a larger class of approaches that combine EEND and VC including, e.g., [6]. This paper discusses a novel clustering approach for EEND-VC.

VC-based diarization approaches first compute speaker embedding vectors for short segments of a recording and then cluster these embeddings to assign speaker labels to each segment. Classical clustering algorithms such as K-means or agglomerative hierarchical clustering (AHC) can be used, but more powerful clustering schemes have been proposed, such as variational Bayesian HMM clustering of x-vector sequences (VBx) [8, 9], which has been widely used in diarization challenges [10–12]. VC approaches can work with an arbitrarily large number of speakers in a recording. However, they assume no speech overlap in a segment, which does not hold for many natural conversations.

EEND [2, 3, 13] is an alternative approach, which uses a neural network to directly output the speech activity for each speaker in a recording even for overlapping regions. EEND can thus handle overlapping speech, making it a competitive alternative to VC. However, it is more challenging to generalize to an arbitrarily large number of speakers [3].



a) HMM of VBx b) HMM of multi-stream VBx

Figure 1: *Schematic diagram of the hidden Markov model (HMM) used for a) the conventional VBx and b) the proposed multi-stream extension ($C = 2$). "A," "B," and "C" represent speakers and ϕ an inactive speaker. P_{loop} is the loop probability. For simplicity, we omitted the states "B A," "C A," "C B".*

EEND-VC has been proposed to combine the strength of both frameworks [5, 6]. It first performs EEND on speech *chunks*[1] using a modified network architecture [5, 6], which estimates the speaker activities and speaker embeddings. Then, it performs VC on the estimated speaker embeddings to stitch together the speaker activities of the same speaker across different chunks. EEND-VC can thus handle overlapping speech as EEND and an arbitrary number of speakers as VC.

Compared to conventional VC, with EEND-VC, each chunk can have multiple speaker embeddings. Therefore, it requires *clustering multi-stream of speaker embeddings* with a constraint that the embeddings from the same chunk should not belong to the same cluster [14]. This has been ensured by using constrained variants of well-known clustering algorithms, such as constrained variants of K-means [15, 16] or constrained agglomerative hierarchical clustering (cAHC) [17]. However, variational Bayesian HMM clustering of x-vector sequences (VBx) [8, 9] has shown better performance than K-means or AHC in diarization and has been widely used in recent challenges [10–12]. In this paper, we investigate the use of VBx for clustering the speaker embeddings of EEND-VC by developing a multi-stream extension of VBx.

VBx is a clustering algorithm based on a Bayesian HMM

The work was partly supported by Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X.

[1]Typically, the chunks of EEND-VC are speech segments longer than those used by VC.

10.21437/Interspeech.2023-628

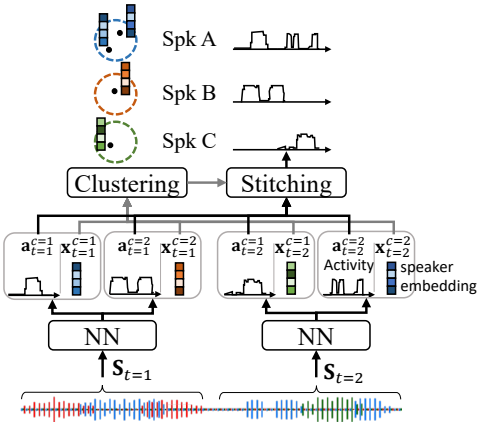


Figure 2: *Schematic diagram of EEND-VC.*

where states are associated with the speakers and transitions between states represent speaker turns, as shown in Fig. 1-a). VBx models the speaker distributions with Gaussians and sets a prior on the parameters of the Gaussians derived from a probabilistic linear discriminant analysis (PLDA) model trained on a large set of speaker embeddings. The parameters of the HMM, the speaker models, and the best assignment of states given a sequence of embeddings are estimated directly on each recording using the variational Bayes (VB) inference. Compared to other clustering approaches, VBx allows modeling the transitions between speakers and offers a principled way of estimating the number of speakers through the VB inference [18]. The formulation of VBx assumes a single speaker for each segment and thus for each embedding.

We propose to generalize the VBx algorithm to handle the multi-stream embeddings of EEND-VC, where each HMM state corresponds to a single or multiple (overlapping) speakers as shown in Fig. 1-b). We call this extension multi-stream VBx (MS-VBx). A naive implementation would exponentially increase the number of parameters of the model and also make the allocation of the speakers difficult. Therefore, we propose to tie together parameters of the HMM states that correspond to the same speaker. This generalization of VBx allows using the state-of-the-art VC approach for EEND-VC.

We show that the proposed MS-VBx naturally implements the constraint required by EEND-VC. Besides, it outperforms cAHC in experiments on the CALLHOME and DIHARD II and III datasets in terms of Diarization error rate (DER) and speaker counting errors.

2. Overview of EEND-VC

Figure 2 shows an EEND-VC system. $\mathbf{S}_t$ are the speech features for the chunk $t = 1, \ldots, T$. EEND-VC consists of a neural network that estimates for each chunk t, the speech activities $\mathbf{a}_t = [\mathbf{a}_t^1, \ldots, \mathbf{a}_t^C] \in [0, 1]^{N \times C}$ and speaker embeddings $\mathbf{x}_t = [\mathbf{x}_t^1, \ldots, \mathbf{x}_t^C] \in \mathbb{R}^{D \times C}$. Here, N represents the number of time frames per chunk, D the dimension of the speaker embedding vectors, and C the number of output streams, which can be a fixed maximum number of active speakers in a chunk [5] or an estimate of the number of active speakers for each chunk [6]. $\mathbf{a}_t^c$ represents the speech activity of the speaker associated with the output stream c, and $\mathbf{x}_t^c$ is the speaker embedding associated with that speaker.

EEND-VC processes each chunk independently. Since the total number of speakers in a recording can exceed the number of speakers in a chunk, and the system can output speakers in an arbitrary order, there is speaker ambiguity at the output of the EEND stage. EEND-VC resolves this ambiguity by clustering the speaker embeddings to associate a global speaker identity to each estimated speech activity, $\mathbf{a}_t^c$. We can then produce the diarization results by stitching the speech activities associated with the same speaker identity.

Prior works have explored various clustering approaches and reported superior performance with cAHC [14]. In this paper, we propose an alternative clustering approach.

3. Proposed multi-stream VBx

As mentioned in section 1, the original derivation of VBx assumes a single embedding vector per segment and no overlap. Consequently, we need to generalize it if we want to use it for clustering the multi-stream embeddings of EEND-VC.

3.1. Multi-stream extension of VBx model

MS-VBx generalizes the VBx algorithm to multi-stream embeddings. With MS-VBx, multiple speakers are associated with an HMM state, and the parameters corresponding to the same speakers are tied together across HMM states, as shown in Figure 1-b). We design the HMM such that the same speaker cannot appear more than once in a given state, i.e., there are no states such as "A A," "B B," or "C C." This design naturally implements the cannot-link constraint required by EEND-VC. In the following, we denote the HMM state index by $s = 1, \ldots, S$, and the speaker index by $g = 1, \ldots, S_g$.

Similarly to VBx, MS-VBx aims at finding the most likely sequence of latent variables $\mathbf{Z} = \{z_1, \ldots, z_T\}$, where z_t defines the hard alignment of the embedding vectors to the HMM states for chunk t. The complete model is expressed as,

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) = p(\mathbf{X}|\mathbf{Z}, \mathbf{Y})p(\mathbf{Z})p(\mathbf{Y}) \quad (1)$$

$$= \prod_t p(\mathbf{x}_t|z_t)p(z_t|z_{t-1}) \prod_g p(\mathbf{y}_g), \quad (2)$$

where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ is the sequence of speaker embeddings, $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{S_g}\}$ is the set of all the speaker-specific latent variables, $\mathbf{y}_g$, and $p(z_t|z_{t-1})$ represents the state transition probability.

The output probability of HMM state s, $p(\mathbf{x}_t|z_t)$, represents the probability that speakers associated with that state are active in that chunk.[2] It is given by,

$$p(\mathbf{x}_t|z_t = s) = \prod_{c=1}^{C} p(\mathbf{x}_t^c|z_t = s), \quad (3)$$

where $p(\mathbf{x}_t^c|z_t = s) = \mathcal{N}(\mathbf{x}_t^c; \mathbf{V}\mathbf{y}_{\tilde{g}}, \mathbf{I})$, and $p(\mathbf{y}_{\tilde{g}}) = \mathcal{N}(\mathbf{y}_{\tilde{g}}; \mathbf{0}, \mathbf{I})$. Here, $\tilde{g}$ is given by $\tilde{g} = \text{Spk}(s, c)$, where $\text{Spk}(s, c)$ is a mapping function that maps the state and stream indexes (s, c) to the speaker index, $\tilde{g}$. In other words, we tie together the Gaussian parameters associated with the same speaker across HMM states and denote by $\mathcal{S}_g$ the set of HMM-sub-states (s, c) associated with the speaker, g. We then have $\text{Spk}(s, c) = g, \forall (s, c) \in \mathcal{S}_g$.

Following the original VBx [9], $\mathbf{y}_g$ is a latent variable acting as a prior on the mean of the speaker models $p(\mathbf{x}_t^c|z_t = s)$, which is derived from a pre-trained PLDA model. $\mathbf{V} = \mathbf{\Phi}^{\frac{1}{2}}$ is a feature transformation matrix, and $\mathbf{\Phi}$ is a diagonal matrix corresponding to the between-speaker covariance matrix in the

[2] We assume here for simplicity that there is an active speaker for each stream, c, unlike what Fig 1-b) suggests. We deal with the possibility of having inactive speakers in Section 3.3.

transformed space of the PLDA model.

As in the original VBx, we set the transition probability to $p(z_t = s|z_{t-1} = s') = (1-P_{\text{loop}})\pi_s + \delta_{s,s'} P_{\text{loop}}$, where P_{loop} is the loop probability, π_s is the probability to transition to state s from the non-emitting node, and δ is the Kronecker delta. P_{loop} is a tuning parameter. We estimate π_s with Bayesian inference.

3.2. Inference

The most likely sequence $\mathbf{Z}$ is obtained using VB inference. This is solved by iteratively updating the approximate posterior distribution $q(\mathbf{Y})$ for a fixed approximate posterior $q(\mathbf{Z})$ and vice versa. The distribution $q(\mathbf{Y}) = \prod_g q(\mathbf{y}_g)$ is updated as,

$$q(\mathbf{y}_g) = \mathcal{N}(\mathbf{y}; \boldsymbol{\alpha}_g, \mathbf{L}_g^{-1}), \tag{4}$$

where

$$\mathbf{L}_g = \mathbf{I} + \frac{F_A}{F_B}\left(\sum_t \sum_{s \in \mathcal{S}_g} \gamma_{t,s}\right)\boldsymbol{\Phi}, \tag{5}$$

$$\boldsymbol{\alpha}_g = \frac{F_A}{F_B}\mathbf{L}_g^{-1} \sum_t \sum_{(s,c) \in \mathcal{S}_g} \gamma_{t,s} \boldsymbol{\rho}_t^c, \tag{6}$$

and $\boldsymbol{\rho}_t^c = \mathbf{V}^\top \mathbf{x}_t^c$. F_A and F_B are scaling factors for the evidence lower bound objective (ELBO) [18].

As for updating $q(\mathbf{Z})$, the state occupancy, $\gamma_{t,s}$, is computed with the forward-backward algorithm using the state output probabilities obtained as,

$$\log \bar{p}(\mathbf{x}_t|s) = F_A \sum_c \left(\boldsymbol{\alpha}_{\tilde{g}} \boldsymbol{\rho}_t^c - \frac{1}{2}\text{tr}\left(\boldsymbol{\Phi}(\mathbf{L}_{\tilde{g}}^{-1} + \boldsymbol{\alpha}_{\tilde{g}}\boldsymbol{\alpha}_{\tilde{g}})\right) - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbf{x}_t^{c\top}\mathbf{x}_t^c \right), \tag{7}$$

where $\tilde{g} = \text{Spk(s,c)}$, and $\text{tr}(\cdot)$ is the trace operator. Note that $\mathbf{L}_g$ is a diagonal matrix because $\boldsymbol{\Phi}$ is diagonal, making the matrix inversion in Eqs. (6) and (7) trivial.

We can estimate π_s as in the conventional VBx [9]. The inference process tends to set π_s to zero for redundant states, dropping these states [9]. With the original VBx, the number of remaining states provides a direct estimate of the number of speakers. For MS-VBx, we also derive the number of speakers from the remaining HMM states, knowing the association between speakers and states.

The inference with MS-VBx is similar to that of the conventional VBx, except for the use of tied Gaussians, which introduces the summation over the set of HMM-sub-states $\mathcal{S}_g$ in Eqs. (5) and (6), and the summation over the different streams in Eq. (7). MS-VBx can thus be easily implemented by extending an existing VBx code.[3]

3.3. Handling inactive speakers

The original VBx removes the silent portions of the signals beforehand since it is challenging to model silent segments with speaker embeddings. In our case, there are chunks where the number of active speakers may be smaller than the number of streams, C. We denote by $\hat{C}_t$ the estimated number of active speakers for chunk t, and by C_s the number of active speakers for the HMM state s.

EEND-VC can provide information about the number of active speakers in a chunk. For example, the EEND with global and local attractors (EEND-GLA) approach [6] has a variable number of outputs corresponding to the number of active speakers in a chunk. In contrast, the EEND-VC scheme [14] has a fixed number of outputs corresponding to the maximum number of speakers in a chunk but is trained to output speech activity close to zero for outputs with no active speakers. We can thus consider that an output, c, is inactive if $\frac{1}{N}\sum_{n=1}^{N} a_{t,n}^c < \tau$, where τ is a predetermined threshold.

We reformulate Eq. (2) by introducing a random variable $\mathbf{W} = [w_1, \ldots, w_T]$ representing the number of speakers in a chunk:

$$p(\mathbf{X}, \mathbf{W}, \mathbf{Z}, \mathbf{Y}) = \prod_t p(\mathbf{x}_t, w_t|z_t) p(z_t|z_{t-1}) \prod_g p(\mathbf{y}_g), \tag{8}$$

where we express the state emission probabilities as,

$$p(\mathbf{x}_t, w_t = C_s|z_t = s) = p(w_t = C_s) \prod_{c=1}^{C_s} p(\mathbf{x}_t^c|z_t = s),$$

and $p(w_t = C_s) = \delta_{\hat{C}_t, C_s}$. We thus rely on the hard decision from EEND-VC to determine w_t.

3.4. Overall procedure

We first train a PLDA model on the speaker embeddings of the EEND-VC model extracted from the training/adaptation datasets. At inference time, we first compute speaker activities and embeddings with EEND-VC for each recording. Then, as with standard VBx, we perform cAHC to obtain a rough estimate of the number of speakers (which VBx can refine/reduce), generate the HMM and initialize the state occupancy. Finally, we run MS-VBx and stitch the activities, $\mathbf{a}_t^c$, based on the clustering results, to obtain the diarization results. Note that compared to standard VBx, MS-VBx requires more HMM states for the same total number of speakers, increasing thus the computational complexity. Future works should address this issue.

4. Related works

Another VB scheme has been proposed recently for clustering of EEND-VC [19]. That work aimed to allow training speaker embedding jointly with the clustering algorithm to reduce the mismatch between training and inference. The scheme relied on using VB-Gaussian mixture model (VB-GMM) such as infinite GMM (iGMM), which are more complex models, as they assume a potentially infinite number of speakers. Besides, with iGMM, it is challenging to implement the cannot-link constraint. In this paper, we base our study on VBx, a simpler and more practical VB scheme.

Unlike [19], in the current stage of our investigations, we apply MS-VBx only at inference time. However, the training and inference mismatch can be compensated using the PLDA model, which transforms the speaker embeddings for VBx. Investigating tighter integration through joint training of EEND-VC with MS-VBx will be part of our future works.

5. Experiments

We evaluate the effectiveness of the proposed method for the CALLHOME [20], DIHARD II [21], and DIHARD III [22] (full set) datasets.

5.1. Settings

Data: The training data comprised 5.5k hours of simulated mixtures created as in [2]. It uses Switchboard-2 (Phase I & II & III), Switchboard Cellular (Part 1 & 2), and the NIST Speaker Recognition Evaluation (2004 & 2005 & 2006 & 2008) with noise from the MUSAN corpus [23] and simulated room im-

[3] https://github.com/BUTSpeechFIT/VBx

pulse responses [24]. The mixtures were of up to 7 speakers, with an average silence duration between utterances of the same speakers of 2 sec ($\beta = 2$).

For the CALLHOME dataset, we used the dev/test set definition of prior works [25], which amounts to 249 and 250 sessions for adaptation and test, respectively.

For each task, we used the development set for adaptation and hyper-parameter tuning and reported results on the test set. We report results using estimated speech activity detection (SAD), corresponding to track 2 of DIHARD II and III challenges. Note that our system directly estimates the speech activity using EEND-VC without using any external SAD.

EEND-VC configuration: We used a EEND-VC model similar to that in [7]. We base our implementation on the publicly available code.[4] It consists of six-stacked Transformer encoder blocks with eight attention heads. We used the pre-trained WavLM-large to obtain the input speech features to the EEND-VC model as in [7]. The input feature dimension was 1024, and the output dimension for each attention block was 256. To produce the final output, the encoder's output is projected with a linear layer into $C = 3$ output streams, where each output consists of the frame-by-frame speaker activity binary decisions and the speaker embedding of 256 dimensions.

We trained the EEND-VC model using chunks of 15 sec and sub-chunks (or subsequences in [6]) of 5 sec, using a similar training scheme as [6]. We trained the model for 70 epochs with a batch size of 2048 and averaged the model over the last five epochs. We used the Adam optimizer with the learning rate scheduler introduced in [26] with 25000 warm-up steps.

For adaptation, we retrained the averaged model on the adaptation set for three epochs. We fixed the WavLM parameters during training but retrained them during adaptation with a learning rate of 10^{-5} and a batch size of one. We performed adaptation using chunks of 30 sec and sub-chunks of 5 sec. At test time, we reduced the sub-chunk length to 1.5 sec. We set the threshold τ to detect the silent speaker at 0.05.

Clustering parameters: We compare the proposed MS-VBx with cAHC [7, 14]. We normalized the embedding with the L2 norm, thus the maximum distance between two embeddings is 2. We reduced the embedding dimension to 32 with a linear discriminant analysis (LDA) model trained on the adaptation set. For cAHC, we used the Euclid distance to cluster the embeddings. We tuned the distance threshold for AHC for values between 1 and 0.8 on the dev set.

For MS-VBx, we trained the PLDA model on the speaker embeddings obtained by processing the adaptation data with EEND-VC and applied the corresponding transformation to the embeddings. We used $F_A = 0.4$, $F_B = 17$ and set the loop probability, P_{loop}, at 0.8. Besides, we applied a median filter on the predicted diarization results with a window of 1.0 sec for CALLHOME and 0.28 sec for DIHARD II and III.

Baseline VC system (SAD + VBx + OSD): We compare our diarization with a VC diarization system, which is based on conventional (i.e., single stream) VBx [9]. First, we used a SAD suited for telephone speech and based on time-delay neural networks and statistical pooling[5] for CALLHOME and a SAD suited for wide-band data released in pyannote [27] for DIHARD II and III. Then, we applied VBx [9] and assigned the second nearest speaker to the segments detected as overlapping with overlapped speech detection (OSD) [28].

[4] https://github.com/nttcslab-sp/EEND-vector-clustering

[5] http://kaldi-asr.org/models/m4

Table 1: *DERs (%) for CALLHOME (CH), DIHARD II (DH II), and DIHARD III (DH III). The numbers in parenthesis indicate the speaker counting performance in terms of ME.*

	System	CH	DH II	DH III
1	SAD + VBx + OSD	13.6	26.8	20.5
2	SAD + VBx + resegm. [29]	-	26.3†	19.3
3	EDA-TS-VAD [30]	11.2	-	-
4	USTC-NELSLIP DH III [11]	-	-	16.8
5	EEND-GLA [6]	11.8	28.3	19.5
6	EEND-VC w/ WavLM [7]	10.4	-	-
7	EEND-VC (cAHC)	11.1 (1.2)	28.2 (3.2)	19.3 (1.3)
8	+ MS-VBx (Proposed)	10.4 (0.6)	26.4 (1.1)	18.2 (0.7)

† results taken from [31].

Evaluation metrics: We evaluated results in terms of DER accounting for the overlap regions with a collar value of 0.25 sec for CALLHOME and 0 sec for DIHARD II and III. We used an estimated number of speakers for all experiments, where the number of speakers was obtained as in [14] for cAHC and as the number of remaining Gaussians with VBx. We also evaluated the speaker counting performance in terms of mean error (ME), $ME = \frac{1}{R}\sum_{r=1}^{R} |C_r - \widehat{C}_r|$, where C_r and $\widehat{C}_r$ are the actual and estimated number of speakers in recording r, and R is the total number of recordings in the test set.

5.2. Results

Table 1 compares the DER for various diarization systems and our proposed EEND-VC with MS-VBx on CALLHOME, DIHARD II and III datasets. The upper part of the table shows the performance of competitive VC-based systems (systems 1 and 2), TS-VAD-based systems (systems 3 and 4), and EEND-VC-based systems (systems 5 and 6). To the best of our knowledge, these systems represent the state-of-the-art for these tasks. System 7 consists of our baseline EEND-VC baseline system using cAHC for clustering. System 8 is the same system using the proposed MS-VBx.

Our baseline EEND-VC (system 7) reproduces the WavLM-based system [7] (system 6). It performs slightly worse on the CALLHOME dataset (i.e., DER of 11.1 vs. 10.4 in [7]), probably because of the difference in the simulated training data. However, it is still a relatively strong baseline, outperforming most prior works on the CALLHOME and DIHARD III datasets. The DIHARD III top system (system 4) performs significantly better, but it is a much more complex system, which combines several diarization approaches [11].

By comparing systems 7 and 8, we confirm that using the proposed MS-VBx reduces DER compared to cAHC for all three datasets. It also significantly reduces speaker counting errors by about 50%. EEND-VC using the proposed MS-VBx achieves similar or superior performance than most prior diarization systems on these tasks. These results demonstrate the potential of the proposed MS-VBx for speaker diarization.

6. Conclusion

In this paper, we have introduced MS-VBx, which is an extension of the VBx algorithm to perform clustering on the multi-stream embeddings generated by recent EEND-VC diarization systems. We have demonstrated the potential of MS-VBx on three popular datasets. In future works, we plan to investigate joint-training of EEND-VC system with the MS-VBx clustering. We will also test the proposed MS-VBx with other EEND-VC frameworks such as EEND-GLA [6].

7. References

[1] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Proc. Interspeech*, 2018, pp. 2808–2812.

[2] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Proc. Interspeech*, 2019, pp. 4300–4304.

[3] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. García, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1493–1507, 2022.

[4] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. Interspeech*, 2020, pp. 274–278.

[5] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7198–7202.

[6] S. Horiguchi, S. Watanabe, P. García, Y. Xue, Y. Takashima, and Y. Kawaguchi, "Towards neural diarization for unlimited numbers of speakers using global and local attractors," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 98–105.

[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[8] M. Diez, L. Burget, S. Wang, J. Rohdin, and J. Černocký, "Bayesian HMM based x-vector clustering for speaker diarization," in *Proc. Interspeech*, 2019, pp. 346–350.

[9] 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*, vol. 71, p. 101254, 2022.

[10] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, A. Silnova, O. Plchot, O. Novotný *et al.*, "BUT system for the second DIHARD speech diarization challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6529–6533.

[11] Y. Wang, M. He, S. Niu, L. Sun, T. Gao, X. Fang, J. Pan, J. Du, and C.-H. Lee, "USTC-NELSLIP system description for DIHARD-III challenge," in *Proc. The Third DIHARD Speech Diarization Challenge Workshop*, 2021.

[12] F. Yu, S. Zhang, P. Guo, Y. Fu, Z. Du, S. Zheng, W. Huang, L. Xie, Z.-H. Tan, D. Wang *et al.*, "Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9156–9160.

[13] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2019, pp. 296–303.

[14] K. Kinoshita, M. Delcroix, and N. Tawara, "Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech," in *Proc. Interspeech*, 2021, pp. 3565–3569.

[15] K. Wagstaff, C. Cardie, S. Rogers, and S. S. Schroedl, "Constrained k-means clustering with background knowledge," in *Proc. 18th International Conference on Machine Learning (ICML)*, 2001.

[16] Y. Yang, T. Rutayisire, C. Lin, T. Li, and F. Teng, "An improved Cop-Kmeans clustering for solving constraint violation based on MapReduce framework," *Fundam. Inf.*, vol. 126, no. 4, p. 301–318, 2013.

[17] I. Davidson and S. S. Ravi, "Using instance-level constraints in agglomerative hierarchical clustering: Theoretical and empirical results," *Data Mining and Knowledge Discovery*, vol. 77, no. 18, pp. 257–282, 2009.

[18] M. Diez, L. Burget, F. Landini, and J. Černocký, "Analysis of speaker diarization based on bayesian HMM with eigenvoice priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 355–368, 2020.

[19] K. Kinoshita, M. Delcroix, and T. Iwata, "Tight integration of neural- and clustering-based diarization through deep unfolding of infinite Gaussian mixture model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8382–8386.

[20] M. Przybocki and A. Martin, *2000 NIST Speaker Recognition Evaluation (LDC2001S97)*. Philadelphia, New Jersey: Linguistic Data Consortium, 2001.

[21] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second DIHARD diarization challenge: Dataset, task, and baselines," in *Proc. Interspeech*, 2019, pp. 978–982.

[22] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The third DIHARD diarization challenge," in *Proc. Interspeech*, 2021, pp. 3570–3574.

[23] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, arXiv:1510.08484.

[24] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220—5224.

[25] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," in *Proc. Interspeech*, 2020, pp. 269–273.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 5998—6008.

[27] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote. audio: neural building blocks for speaker diarization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7124–7128.

[28] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007, pp. 683–686.

[29] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech*, 2021, pp. 3111–3115.

[30] D. Wang, X. Xiao, N. Kanda, T. Yoshioka, and J. Wu, "Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization," *arXiv preprint arXiv:2208.13085*, 2022.

[31] S. Horiguchi, S. Watanabe, P. García, Y. Takashima, and Y. Kawaguchi, "Online neural diarization of unlimited numbers of speakers using global and local attractors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 706–720, 2023.