Breaking the Pair: Evaluating dyadic interaction via speaker switching

Nishchay Nilabh, Neeraj Kumar Sharma
Mehta Family School of Data Science and Artificial Intelligence
Indian Institute of Technology Guwahati, 781039, India
n.nishchay@alumni.iitg.ac.in, neerajs@iitg.ac.in

***Abstract*—Speakers in dialogue continuously adapt their communicative behavior across acoustic, lexical, and semantic dimensions, a phenomenon known as conversational entrainment. Modeling this process requires representations that capture the global structure of interaction, yet prior approaches fail to disentangle dyad-specific patterns from speaker-specific traits, limiting their ability to capture true conversational adaptation. We address this with the Dyadic Distance Matrix (DDM), which encodes all pairwise distances between the turns of two speakers over an entire conversation, capturing long-range cross-speaker dependencies. This raises a key question: does the DDM represent genuine interaction, or merely reflect individual speaker characteristics? We propose the speaker-switch test, a principled control in which one speaker's turns are replaced with those from an unrelated speaker drawn from a different conversation. This preserves turn-level statistics while disrupting the original dyadic co-adaptation. The ability to distinguish real from switched DDMs thus directly evaluates whether the representation encodes interaction-specific structure. Across four embedding types and four classifiers (MLP, CNN, ResNet-50 and ViT) on the CANDOR corpus, real DDMs are consistently distinguishable from their switched counterparts, and this discriminability is robust to a fully speaker-disjoint data split. Comparisons with synthetic conversations created using LibriSpeech show higher discriminability in read speech, highlighting the role of prosodic variability in naturalistic conversations. Grad-CAM analysis further reveals distinct structural signatures driving classification. These results establish the speaker-switch test as a robust diagnostic for validating representations of dyadic conversational interaction.**

***Index Terms*—conversational entrainment, dyadic distance matrix, speaker-switch, Grad-CAM, spoken dialogue**

I. Introduction

Speakers in dialogue continuously adapt their communicative behavior to their interlocutors across acoustic, lexical, and semantic dimensions [1]–[3]. This phenomenon, known as *conversational entrainment*, is formalized under Communication Accommodation Theory (CAT) [4], [5], which predicts both convergence and divergence depending on communicative goals. Capturing entrainment computationally therefore requires representations that move beyond local, adjacent-turn similarity to model the evolving global structure of an entire conversation. However, most prior computational approaches rely on local turn-level metrics [6], [7], which conflate two distinct sources of structure: (i) *conversation-specific* patterns arising from the joint behavior of a dyad, and (ii) *speaker-specific* characteristics intrinsic to each participant. As a result, models may rely on speaker identity rather than true dyadic co-adaptation, limiting their validity as measures of conversational entrainment.

To address this limitation, we use the Dyadic Distance Matrix (DDM), a representation that encodes all pairwise cosine distances between the turn embeddings of two speakers across an entire conversation. By capturing long-range cross-speaker dependencies and global temporal structure, the DDM provides a compact representation of interaction dynamics. This naturally raises a fundamental question: does the DDM encode genuine *interaction*, or does it primarily reflect the statistical properties of individual speakers?

We answer this question through the *speaker-switch test* (Fig. 1), a principled control designed to disrupt dyadic interaction while preserving speaker-level properties. Given a conversation between speakers A and B, we construct a synthetic counterpart by replacing B's turn sequence with turns from an unrelated speaker drawn from a different conversation. This produces a controlled contrast (e.g., CANDOR-Real vs. CANDOR-Switch) that maintains individual feature distributions and turn structure, while eliminating genuine co-adaptation between the original pair. A model that reliably distinguishes real from switched DDMs therefore provides direct evidence that the representation encodes interaction-specific structure. Our contributions are as follows:

1) We introduce the speaker-switch test, a principled evaluation framework for distinguishing interaction-specific structure from speaker-specific effects in conversational representations.
2) We conduct a systematic evaluation across four embedding types spanning acoustic and semantic layers, comparing ResNet-50, CNN, MLP, and ViT classifiers.
3) We perform a cross-corpus analysis using the LibriSpeech read-speech dataset [8] to examine how prosodic variability in naturalistic speech affects acoustic discriminability.
4) We provide a Grad-CAM-based [9] interpretability analysis identifying which structural regions of the DDM carry discriminative signal.

II. Background

A. Dyadic Distance Matrix

We represent a two-speaker conversation using the Dyadic Distance Matrix (DDM), which captures cross-speaker similarity

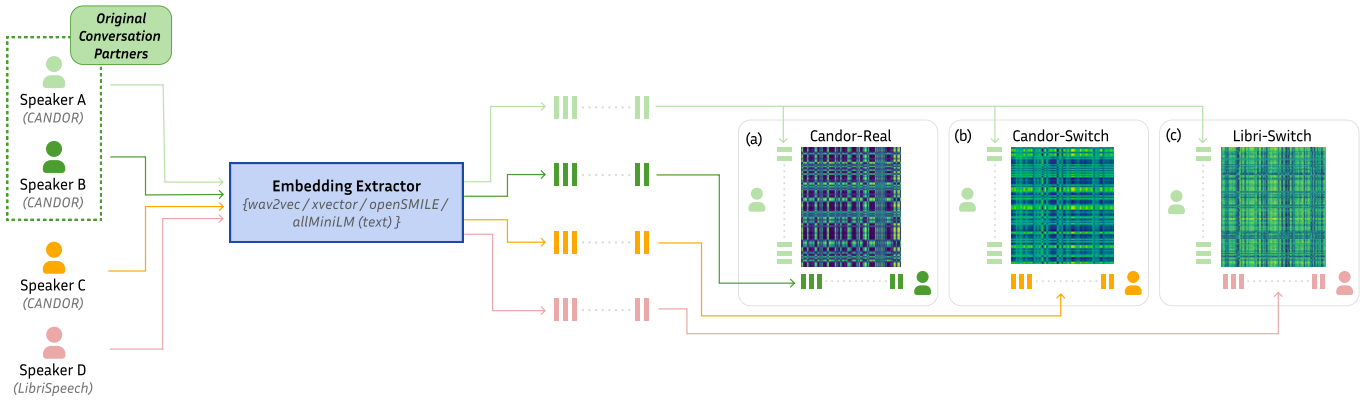

Fig. 1. Dyadic Distance Matrix (DDM) generation pipeline. Turn-level embeddings (a_i, b_j) are extracted using acoustic or semantic models to compute the global cross-speaker distance matrix (M). The schematic illustrates the formulation of the three experimental pairings: (a) genuine interaction between original conversation partners (CANDOR-Real), (b) within-corpus control using an unrelated interlocutor (CANDOR-Switch), and (c) the cross-corpus read-speech (Libri-Switch) using Speaker D's speech from the LibriSpeech dataset [8].

across all turns. Let $\mathbf{a}_i$ and $\mathbf{b}_j$ denote the embedding sequences corresponding to the turns of speakers A and B, respectively. The DDM is defined as a matrix $M \in \mathbb{R}^{|\mathbf{a}|\times|\mathbf{b}|}$, where each entry (i, j) encodes the cosine distance between turns i and j:

$$M_{i,j} = 1 - \frac{\mathbf{a}_i^\top \mathbf{b}_j}{\|\mathbf{a}_i\|_2 \, \|\mathbf{b}_j\|_2}. \tag{1}$$

Each entry of M reflects the distance (dissimilarity) between a pair of turns across speakers, providing a global view of interaction over the entire conversation. Unlike adjacent-turn metrics [6], [7], which focus on local dependencies, the DDM encodes all cross-speaker turn combinations, enabling the analysis of long-range interactional structure.

Following Communication Accommodation Theory (CAT), we compute DDMs across three complementary representational layers: (i) an *acoustic layer* (wav2vec 2.0, openSMILE, x-vector), capturing prosodic, paralinguistic and speaker-specific cues; and (ii) a *semantic layer* (all-MiniLM), capturing lexical and topical alignment between speakers.

III. Methodology

A. Datasets

We use the CANDOR corpus [10], comprising 1,656 spontaneous dyadic conversations (approximately 850 hours) from over 1,400 speakers recorded via video-calling platforms. Each session lasts approximately 31 minutes and includes word-level timestamps, speaker diarization, and transcripts. We apply Cliffhanger sentence-level segmentation to obtain semantically coherent speaker turns, and exclude sessions with fewer than 20 turns per speaker.

To assess the role of prosodic variability, we conduct a complementary cross-corpus experiment using LibriSpeech [8]. As a large-scale read-speech corpus, LibriSpeech provides a controlled setting with relatively uniform prosody, in contrast to the highly variable naturalistic speech in CANDOR.

B. Task Formulation

We cast the speaker-switch test as a binary classification task over DDMs. Given a Dyadic Distance Matrix M, the objective is to predict whether it is derived from a *real* conversation (CANDOR-Real) or a *speaker-switched* counterpart (CANDOR-Switch).

Formally, the classifier learns a mapping $f : M \to \{0, 1\}$, where the label indicates the presence or absence of genuine dyadic interaction. Successful discrimination implies that the DDM encodes interaction-specific structure beyond speaker-level statistics.

C. Speaker-Switch DDM Construction

Given an original conversation $\mathcal{C}_1 = (A_1, B_1)$ and an unrelated conversation $\mathcal{C}_2 = (A_2, B_2)$, the switched DDM for $\mathcal{C}_1$ is constructed by replacing B_1 with B_2, while retaining A_1's turn sequence. Conversation pairs are matched by approximate turn count to preserve matrix dimensions.

$$M^{\text{sw}}_{i,j} = 1 - \frac{{\mathbf{a}_i^{(1)}}^\top \mathbf{b}_j^{(2)}}{\|\mathbf{a}_i^{(1)}\|_2 \, \|\mathbf{b}_j^{(2)}\|_2}. \tag{2}$$

This construction preserves: (i) each speaker's embedding distribution; (ii) within-conversation temporal structure; and (iii) global pairwise distance statistics, while eliminating (iv) genuine dyadic co-adaptation between A_1 and B_1.

D. Feature Extraction

We evaluate four embedding types corresponding to the three CAT representational layers:

- **wav2vec 2.0** [11]: 768-dimensional representations obtained via mean-pooling of the final transformer layer; the base pre-trained model is used to retain prosodic information.
- **x-vector** [12]: 512-dimensional speaker embeddings extracted from a pre-trained ECAPA-TDNN model via `pyannote.audio` [13].

- **openSMILE (eGeMAPS)** [14], [15]: 88-dimensional acoustic descriptors capturing pitch, energy, spectral, and cepstral features.
- **all-MiniLM** [16]: 384-dimensional sentence embeddings encoding contextual semantic information.

Per-speaker z-normalization is applied to wav2vec 2.0, x-vector, and openSMILE embeddings to remove static speaker-level biases prior to DDM computation.

E. Classifiers

We evaluate four classifier architectures. Because each DDM is a two-dimensional image-like array whose discriminative cues are spatially distributed (Sec. IV-C), our primary model is a ResNet-50 backbone [17] (in PyTorch [18]), whose hierarchical residual features capture multi-scale spatial structure; its final fully connected layer is replaced by a dropout layer ($p = 0.3$) and a sigmoid output. A three-layer CNN and a multilayer perceptron (MLP) serve as shallow baselines, and a ViT-B/16 [19] (ImageNet-pretrained, 16×16 patches) provides a globally aware, attention-based comparison.

All models use the Adam optimizer (learning rate 1×10^{-4}, batch size 32) with early stopping on validation loss. The data are split 70/15/15 (train/validation/test) stratified by conversation; DDMs are resized via bilinear interpolation to 64×64 (224×224 for ViT) and standardized to unit variance. We report accuracy, macro-F1, and Equal Error Rate (EER), the threshold-independent operating point at which false-acceptance and false-rejection rates coincide. To rule out speaker-identity leakage, we additionally re-evaluate under a fully speaker-disjoint split (no speaker shared across train/validation/test): discriminability is essentially unchanged, with accuracy, macro-F1, and EER all remaining within a small margin of the standard-split values across all modalities, and a probe stratifying test pairs by whether their speakers appear in training performs near-identically on seen and unseen speakers. This confirms the classifiers capture interaction structure rather than speaker identity.

F. Cross-Corpus Evaluation

To isolate the effect of prosodic variability, we construct cross-corpus switched DDMs by replacing one speaker in a CANDOR conversation with a randomly sampled speaker from LibriSpeech. Specifically, given a CANDOR conversation $\mathcal{C}_1 = (A_1, B_1)$, we retain A_1's turn sequence and replace B_1 with speech segments from an unrelated LibriSpeech speaker, matched by approximate turn count.

This results in a synthetic interaction between naturalistic conversational speech and read speech, preserving the structure of one side of the dialogue while introducing a controlled mismatch in speaking style. Applying the same classification pipeline allows us to assess how reduced prosodic variability in read speech affects discriminability relative to fully naturalistic dyadic interactions.

G. Grad-CAM Interpretability

We apply Gradient-weighted Class Activation Mapping (Grad-CAM) [9] to the trained ResNet-50 to identify discriminative regions within the DDM. For each correctly classified test sample, Grad-CAM heatmaps are computed and averaged across samples within each condition and embedding type. These aggregated maps reveal the structural loci of interaction-specific information captured by the model.

IV. Results and Discussion

A. Classification: Real vs. Switch

Table I presents classification results across all modalities, model architectures, and corpora. On CANDOR, the semantic embedding (all-MiniLM) achieves perfect discrimination with ResNet-50 (Acc=1.000, EER=0.000), and remains highly discriminable even with the shallow CNN (Acc=0.952) and MLP (Acc=0.857). Semantic alignment is thus the most easily detected signal, robust to classifier choice. However, as we discuss below, this discriminability is partly attributable to topical divergence between unrelated conversations rather than to fine-grained co-adaptation alone.

For acoustic embeddings on CANDOR, ResNet-50 consistently outperforms the shallower architectures. x-vector ResNet-50 achieves Acc=0.682 versus CNN at 0.597, while the MLP collapses to majority-class prediction (Acc=0.500, F1=0.000, i.e. it assigns all samples to a single class), unable to learn the interaction patterns encoded in the structural DDM. A similar collapse occurs for openSMILE MLP (Acc=0.467, below chance), whereas ResNet-50 attains 0.692. Notably, the ViT, despite its global self-attention, does not surpass ResNet-50 on CANDOR (x-vector 0.658 vs. 0.682; openSMILE 0.554 vs. 0.692), exceeding it only on wav2vec 2.0 (0.647 vs. 0.620). This indicates that hierarchical, locally aggregated spatial features, rather than global attention alone, best capture the distributed interaction structure, motivating ResNet-50 as our primary model.

That all four embedding types exceed chance on CANDOR with at least one model provides validation: DDMs encode genuine interaction-specific structure that cannot be recovered by combining arbitrary speakers. Semantic representations are the most discriminable, but this result is partly trivial: unrelated conversations differ in topic, so a switched semantic DDM loses topical continuity by construction. The near-perfect all-MiniLM score should therefore be read as an upper bound rather than as evidence of fine-grained semantic co-adaptation; the more informative evidence of interaction-specific structure comes from the acoustic embeddings, where topic is not a confounding cue. A stronger semantic control that pairs topic-matched conversations before switching is left to future work.

B. Cross-Corpus Results: The Role of Prosodic Variability

The contrast between CANDOR and LibriSpeech results is the most striking finding in Table I. For x-vector embeddings, all four classifiers achieve perfect accuracy on LibriSpeech (Acc=1.000, EER=0.000), including the MLP that collapses on CANDOR. LibriSpeech consists of read speech where intonation is controlled and within-speaker variability is low. In this setting, x-vector embeddings robustly encode stable,

TABLE I
REAL VS. SWITCH CLASSIFICATION ACROSS MODELS AND MODALITIES. BEST PER METRIC/MODALITY/CORPUS IN **BOLD**.

		CANDOR-Switch				Libri-Switch			
Modality	**Metric**	**MLP**	**CNN**	**ResNet-50**	**ViT**	**MLP**	**CNN**	**ResNet-50**	**ViT**
wav2vec 2.0	Acc	0.584	0.552	0.620	**0.647**	**1.000**	0.823	0.904	0.918
	F1	0.559	0.433	0.649	**0.651**	**1.000**	0.820	0.904	0.919
	EER	0.402	0.474	0.394	**0.353**	**0.000**	0.169	0.092	0.080
x-vector	Acc	0.500	0.597	**0.682**	0.658	**1.000**	**1.000**	**1.000**	**1.000**
	F1	0.000	0.572	**0.687**	0.663	**1.000**	**1.000**	**1.000**	**1.000**
	EER	0.485	0.388	**0.315**	0.339	**0.000**	**0.000**	**0.000**	**0.000**
openSMILE	Acc	0.467	0.575	**0.692**	0.554	**0.975**	0.842	0.908	0.942
	F1	0.458	0.557	**0.722**	0.550	**0.974**	0.832	0.904	0.940
	EER	0.533	0.450	**0.350**	0.467	**0.000**	0.150	0.083	0.067
all-MiniLM	Acc	0.857	0.952	**1.000**	0.744	**1.000**	0.994	0.998	0.998
	F1	0.858	0.950	**1.000**	0.739	**1.000**	0.994	0.998	0.998
	EER	0.149	0.064	**0.000**	0.211	**0.000**	0.004	0.004	0.004

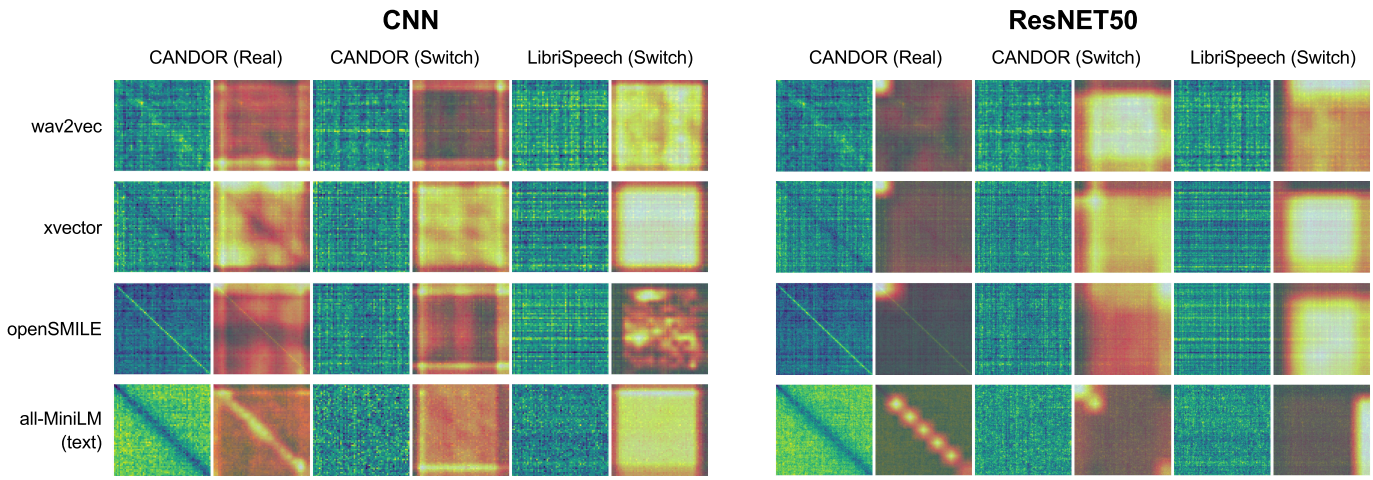

Fig. 2. Mean Grad-CAM heatmaps for ResNet-50 and CNN models across all embedding types (columns: wav2vec 2.0, x-vector, openSMILE, all-MiniLM). Rows show averaged input DDMs and Grad-CAM activations for real and switch conditions, where warmer colors indicate higher activation.

distinct speaker signatures, making cross-speaker pairing trivially detectable regardless of classifier capacity.

On CANDOR, speakers exhibit rich intonation variation, voice modulation, and prosodic accommodation. Post-z-normalization, these residual patterns of within-speaker variability and cross-speaker adaptation reduce the discriminability of switched DDMs, and only ResNet-50 retains meaningful performance. The acoustic embedding results similarly follow this pattern: wav2vec 2.0 and openSMILE both show markedly higher accuracy on LibriSpeech than on CANDOR across all models. The LibriSpeech result thus serves as an important sanity check: the difficulty of distinguishing real from switched acoustic DDMs on CANDOR is not a failure of the representation, but rather a reflection of the genuine complexity of conversational acoustic adaptation.

Notably, the MLP achieves Acc=1.000 on wav2vec 2.0 LibriSpeech despite underperforming ResNet-50 on CANDOR. This inversion suggests that read speech embeddings contain a simple low-dimensional discriminative feature exploitable by a linear-style classifier, while conversational speech requires multi-scale spatial reasoning to separate real from switched patterns.

C. Grad-CAM Analysis

Fig. 2 shows mean Grad-CAM activations for ResNet-50 and CNN, aggregated over correctly classified test samples under real and switch conditions.

Semantic modality: For all-MiniLM, ResNet-50 Grad-CAM maps under the real condition exhibit strong activation concentrated along and near the DDM diagonal. This diagonal structure reflects temporal proximity: turns occurring close in time are semantically similar, consistent with topic progression

and shared reference in natural conversation. Under the switch condition, the diagonal activation disappears, replaced by diffuse or off-diagonal patterns, confirming that the diagonal is a signature of genuine conversational content alignment rather than any individual speaker property.

Acoustic modality: For acoustic embeddings, ResNet-50 activates more globally rather than concentrating on the diagonal. This is consistent with the literature documenting acoustic accommodation operating over longer time scales [1]: prosodic adaptation is distributed across the full conversation timeline rather than localized to adjacent turns. Notably, Grad-CAM visualizations for ResNET-50 real CANDOR DDMs display a localized concentration in the top-left quadrant; while this pattern warrants further investigation, the overall activation remains broadly distributed across the matrix.

ResNet-50 vs. CNN: Within Fig. 2, the CNN exhibits broader and less spatially structured activations than ResNet-50, particularly for acoustic modality. This explains the performance gap on CANDOR: the hierarchical feature extraction of ResNet-50 enables detection of complex, spatially distributed interaction signatures, whereas the three-layer CNN lacks sufficient depth.

V. Conclusion

We have presented a systematic speaker-switch validation of Dyadic Distance Matrices as representations of conversational interaction. Across four embedding types on CANDOR, switched DDMs are consistently distinguishable from real DDMs, validating that the DDM encodes genuine dyad-specific interaction structure beyond individual speaker characteristics. The LibriSpeech cross-corpus experiment shows that the lower discriminability of acoustic DDMs on CANDOR reflects the genuine prosodic variability of naturalistic conversation rather than a weakness of the representation. Grad-CAM analysis shows the structural signatures in the DDMs that drive classification: diagonal temporal alignment for semantic features and distributed global structure for acoustic features. These findings provide an interpretable foundation for using DDMs as a representation of conversational dynamics. Future work will integrate these validated DDM representations into predictive architectures. By isolating co-adaptation from speaker-specific traits, DDMs may enable applications such as outcome prediction, mutual understanding assessment, and more responsive dialogue systems.

Acknowledgment

The authors thank Srikanth Raj Chetupalli and Shreyas Ramoji for discussions on this work.

References

[1] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. Interspeech*, 2011, pp. 3081–3084.

[2] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 22, no. 6, pp. 1482–1493, 1996.

[3] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proc. NAACL-HLT*, 2012, pp. 11–19.

[4] H. Giles, D. M. Taylor, and R. Bourhis, "Towards a theory of interpersonal accommodation through language: Some Canadian data," *Language in Society*, vol. 2, no. 2, pp. 177–192, 1973.

[5] N. Coupland, H. Giles, and J. M. Wiemann, "Accommodation theory: Communication, context, and consequence," in *Contexts of Accommodation: Developments in Applied Sociolinguistics*, H. Giles, J. Coupland, and N. Coupland, Eds. Cambridge University Press, 1991, pp. 1–68.

[6] M. Nasir, B. Baucom, C. Lord, S. Narayanan, and P. Georgiou, "Modeling vocal entrainment in conversational speech using deep unsupervised learning," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1568–1579, 2022.

[7] R. Lahiri, M. Nasir, C. Lord, S. H. Kim, and S. Narayanan, "A context-aware computational approach for measuring vocal entrainment in dyadic conversations," in *Proc. ICASSP*, 2023.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE/CVF ICCV*, 2017, pp. 618–626.

[10] A. Reece, G. Cooney, P. Bull, C. Chung, B. Dawson, C. Fitzpatrick, T. Glazer, D. Knox, A. Liebscher, and S. Marin, "The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation," *Science Advances*, vol. 9, no. 13, p. eadf3197, 2023.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[12] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[13] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe," in *Proc. Interspeech*, 2023, pp. 1983–1987.

[14] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*, 2010, pp. 1459–1462.

[15] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF CVPR*, 2016, pp. 770–778.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.