Investigating the Effects of Large-Scale Pseudo-Stereo Data and Different Speech Foundation Model on Dialogue Generative Spoken Language Model

Yu-Kuan Fu[1,2], *Cheng-Kuang Lee*[1], *Hsiu-Hsuan Wang*[2], *Hung-yi Lee*[2]

[1]NVIDIA
[2]National Taiwan University

r11942083@ntu.edu.tw, ckl@nvidia.com, b09902033@ntu.edu.tw, tlkagkb93901106@gmail.com

Abstract

Recent efforts in Spoken Dialogue Modeling aim to synthesize spoken dialogue without the need for direct transcription, thereby preserving the wealth of non-textual information inherent in speech. However, this approach faces a challenge when speakers talk simultaneously, requiring stereo dialogue data with speakers recorded on separate channels—a notably scarce resource. To address this, we have developed an innovative pipeline capable of transforming single-channel dialogue data into pseudo-stereo data. This expanded our training dataset from a mere 2,000 to an impressive 17,600 hours, significantly enriching the diversity and quality of the training examples available. The inclusion of this pseudo-stereo data has proven to be effective in improving the performance of spoken dialogue language models. Additionally, we explored the use of discrete units of different speech foundation models for spoken dialogue generation.

Index Terms: spoken dialogue modeling, speech generation, human-computer interaction

1. Introduction

Spoken dialogue, characterized by spontaneous turn transitions and occasional overlaps, embodies the natural flow of human communication [1]. These moments of silence, laughter, and overlapping speech are vital cues within conversations [2, 3, 4]. Yet, current dialogue systems like Siri often overlook these signals, relying on a conventional pipeline of transcribing, generating textual response, and converting it to speech, resulting in unnatural dialogues [5].

現在熱門議題：讓模型辨認文字以外的聲音

現在已經能夠直接將語音訊號編碼為離散token，不需隱含地轉錄成文字，這種方式可以同時保留語言與非語言的線索

Advancements in self-supervised (SSL) models [6, 7, 8, 9] and textless spoken language modeling [10, 11, 12] now enable encoding speech signals directly into discrete tokens without implicitly transcribing to text. This preserves both verbal and nonverbal cues, aligning with human turn-taking behaviors. The introduction of the dialogue generative spoken language model (dGSLM) [13] marks a milestone, employing a dual tower transformer to process input tokens from separate channels and enhance dialogue generation. While dGSLM has shown to generate natural human dialogue, it faces challenges in generating semantic coherence dialogue due to limited data. The training data of dGSLM is a notably scarce resource because it requires stereo dialogue data.

困境：dGSLM需要立體聲對話資料，訓練資料稀少

Acquiring stereo dialogue data is challenging. In contrast, single-channel dialogue content is abundant. For example, in this paper, we gathered approximately 20k hours of podcasts. While the speakers of podcasts are mixed into one channel, which dGSLM can not directly use, we proposed a pipeline to automatically generate pseudo-stereo data through the following process: first, employing speaker diarization to pinpoint segments featuring two speakers; next, leveraging source separation techniques to isolate overlapping frames; and finally, applying speaker verification to allocate the separated speech to its respective speakers. This meticulous process yielded a substantial 15.6k hours of pseudo-stereo dialogue training data[1].

我們提出了可以自動產生擬立體聲（pseudo-stereo）資料的方法

Additionally, we explored the use of discrete units from state-of-the-art (SOTA) foundation models for dGSLM. Our investigation revealed that merely scaling foundation models led to poor vocoder performance when resynthesizing speech from discrete units. However, employing an ASR fine-tuned foundation model showed significant improvements across all aspects. By integrating the ASR fine-tuned model with our pseudo-stereo data, dGSLM excelled in producing dialogue semantic coherence[2].

2. Related Work

2.1. Speech Unit Language Modeling

The Textless NLP series of works introduced a framework to address speech tasks using discrete NLP approaches. This framework can be broadly divided into three components [10, 12]: encoding speech into discrete units by clustering its pretrained self-supervised learning (SSL) representations [7, 8, 6, 9], autoregressively generating discrete units, and resynthesizing speech from these discrete units[14].

Quantized unit sequences are shorter than the original signal, significantly reducing computational costs and enabling the modeling of speech through NLP approaches. Additionally, this framework has successfully generated natural speech without textual information [10]. With more fine-grained discrete units, it can preserve non-verbal vocalizations and even control the prosody of speech [11].

2.2. Spoken Dialogue Generation

Current commercial spoken dialogue generation models (e.g., Siri, Alexa, Google Assistant) divide spoken dialogue into three components: ASR, text-base dialogue language model, and TTS [5]. These models primarily focus on semantic content, often neglecting other speech related information, resulting in unnatural dialogue.

Recent work has focused on generating a diverse set of spoken dialogues that incorporate turn-taking and acoustic in-

[1]We open source our dataset: https://huggingface.co/datasets/YuKuanFu/Podcast_Dialogue

[2]Generation samples can be found at https://anonymous78264.github.io/pseudo-stereo-data/?fbclid=IwAR0MGdFnQeUcnojhQGGk0HaAYgBxhnhblIpU3xnGRNfPPO_hxHOf6Ea_PGM

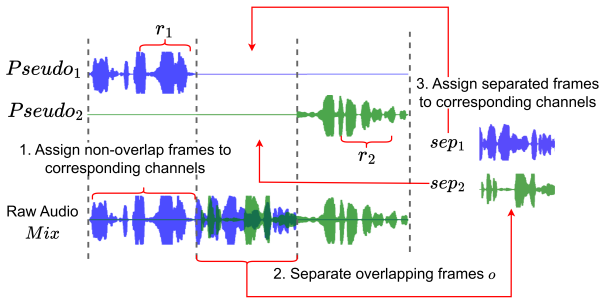

Figure 1: *The pipeline of generating pseudo-stereo data from single-channel dialogue data. We split the process into 3 steps: speaker diarization, source separation, and speaker verification.*

formation to mimic human-like conversational flow [15, 16]. dGSLM [13] is the first spoken dialogue generation model that can successfully generated natural dialogue with non-verbal signals. It follows the three components mentioned in Sec.2.1. It first encodes two speakers by clustering the representations of SSL models separately, and generates two parallel discrete unit sequences from a dual-tower transformer language model. Moreover, it splits the next unit prediction into edge-unit prediction and duration prediction during training for better performance. Although dGSLM can generate natural dialogue, the lack of data results in poor performance in semantic coherence within dialogues.

3. Method

Our methodology is an enhancement of the dGSLM framework [13]. dGSLM decomposes spoken dialogue generation into three components: the Speech-to-Units encoder, the dual-tower unit dialogue language model, and the Units-to-Speech vocoder. The original dGSLM model encountered difficulties in producing semantically coherent dialogue, a problem attributed to insufficient data and suboptimal unit encoding [13]. Thus in this study, we places emphasis on two main objectives: augmenting the training of dGSLM with additional data and improving the performance of unit encoding.

The following sections present a detailed overview of our approach, covering the process of generating pseudo-stereo data from single-channel recording and the Speech-to-Units encoders used.

3.1. Pseudo-stereo Data Generation

dGSLM utilized the Fisher Telephone conversation collection protocol [17] as the training dataset for its components. The Fisher dataset is a well-known benchmark for spoken dialogue datasets, containing around 2000 hours of telephonic dialogues. Each conversation in the Fisher dataset involves two speakers recorded in separate audio channels, resulting in a stereo audio setup.

However, most available dialogue data do not separate speakers, often existing as single-channel audio. To overcome this limitation, we developed a pipeline to disentangle speakers from single-channel audio and store their speech into separate channels, creating pseudo-stereo data. This strategy enhanced the training dataset for dGSLM by incorporating a wider variety of data, which enriched the model's learning process and boosted its overall performance. The framework of this approach is illustrated in Fig. 1. This process is divided into three steps:

3.1.1. Speaker Diarization

We used a Speaker Diarization model (SD) [18, 19] to disentangle speakers from a segment of raw audio (Mix), resulting in a set of speaker-duration tuples:

$$SD(Mix) = \{(p, s, e) \mid p \in \{p_1, p_2\},\ e > s\} \quad (1)$$

Here, we only preserved segments that contains exactly two speakers. p_1 and p_2 represent the first and second speakers, while s and e are the start and end frames of the segment that p is speaking.

We created a 2-channel audio ($Psuedo$) filled with silences, matching the length of Mix (we use $Psuedo_1$, $Psuedo_2$ to indicate its first, second channel respectively). With the output of SD, we identified non-overlapping frames (O') and overlapping frames (O):

$$\begin{aligned} O' &= \{(p, s, e) \mid \text{only p is speaking from } s \text{ to } e\} \\ O &= \{(s, e) \mid p_1, p_2 \text{ are both speaking from } s \text{ to } e\} \end{aligned} \quad (2)$$

We then assigned non-overlapping intervals in Mix to $Pseudo$ accoring to O' (correspond to the step 1 in Fig. 1):

for i in 1, 2; **do**
  for (p_i, s, e) in O'; **do**
    $Pseudo_i[s:e] \leftarrow \text{Mix}[s:e]$
  end for
end for

3.1.2. Source Separation

Source separation model (SS) [20, 21, 22] can effectively separate the overlapping speech (o):

$$SS(o) = sep_1, sep_2 \quad (3)$$

Here, sep_1 and sep_2 represent the separated speech for the two speakers (correspond to Step 2 Fig. 1). However, at this juncture, it remains unclear which of these segments can be attributed to speaker p_1 and which to speaker p_2.

Note that we can not directly apply SS on entire Mix, as current source separation algorithms struggle with sparse overlapping speech [23].

3.1.3. Speaker Verification

We match sep_1 and sep_2 their corresponding speaker by speaker verification (SV) [24, 25]. SV compares two audio clips and returns the similarity of their speaker embeddings (correspond to the step 3 in Fig.1):

$$SV(r_i, sep_j) = Sim_{i,j} \quad (4)$$

Here, r_i is a reference clip randomly selected from non-overlapping frames from p_i (see Fig.1). We assigned sep_1 and sep_2 based on the similarities:

for (s, e) in O; **do**
  $sep_1, sep_2 = SS(Mix[s:e])$
  $Sim_{i,j} = SV(r_i, sep_j)$ for i, j in $(1, 2) \times (1, 2)$
  if $Sim_{1,1} + Sim_{2,2} > Sim_{1,2} + Sim_{2,1}$ **then**
    $Pseudo_1[s:e], Pseudo_2[s:e] \leftarrow sep_1, sep_2$
  else
    $Pseudo_1[s:e], Pseudo_2[s:e] \leftarrow sep_2, sep_1$

8: **end if**
9: **end for**

By employing this pipeline, we successfully disentangled speakers in spoken dialogue, enabling scalable creation of diverse and large dialogue audio corpora.

3.2. Discrete Units Encoder

The original dGSLM utilized HuBERT [7] pretrained on the Fisher dataset to obtain better phonetic discriminated representations, followed by training a k-means model to cluster its hidden representation. However, pretraining HuBERT on target domain data is time-consuming and its generalizability is uncertain. To investigate the suitability of existing publicly available speech foundation models for spoken dialogue language model, we explored the base and large versions of HuBERT and WavLM [8], which are state-of-the-art speech foundation models.

Additionally, we incorporated HuBERT large fine-tuned (HuBERT large ft) as the speech encoder. HuBERT large ft is obtained by fine-tuning HuBERT large by connectionist temporal classification (CTC) objective. ASR models tend to provide more phonetic information, which might be beneficial for the spoken dialogue language modeling task[3].

4. Experiment

4.1. Data Setup

In addition to Fisher Dataset [17] used in dGSLM [13], we scrape about 20k hours of podcast raw audio from Apple Podcast[4], and after applying our pipeline described in Sec. 3.1, we create about 15.6k hours of pseudo-stereo spoken dialogue data.

To our knowledge, the existing large-scale datasets containing spoken dialogue are GigaSpeech [26] and Spotify Podcast Dataset [27]. However, GigaSpeech contains at most 7500 hours of raw dialogue data from Podcast and YouTube, which does not increase the dataset too much (after filtering). While Spotify Podcast Dataset comprises about 200k hours of podcast data, Spotify ceased sharing this corpus, prompting us to scrape the podcast data by ourselves.

For validation and evaluation purposes, we randomly chose 1 conversation for each speaker pair in the Fisher, respectively, and we randomly sampled 1% of the pseudo-stereo data from uniform channels for validation and evaluation. This subset represents a balanced selection for a fair comparison between models.

4.2. Pseudo-Stereo Data Pipeline

Our approach to creating pseudo-stereo data involved several steps, utilizing a combination of tools and models for effective speaker diarization, source separation, and speaker verification.

For speaker diarization, we adopted the pyannote toolkit[5], we set the size of segments from 30 seconds to 120 seconds depending on how long the two speakers are speaking.

To separate overlapping speech segments and disentangle the speakers, we trained a Sepformer model [20] on the 16k max split of the Libri2Mix dataset [23] using the speechbrain toolkit [28]. Sepformer model is a SOTA transformer-based model for source separation tasks, providing high-quality separation of overlapping speech.

For speaker verification, we utilized an open-source ECAPA-TDNN checkpoint[6] [24, 28]. ECAPA-TDNN model is a deep neural network architecture designed for speaker recognition tasks, offering robust speaker embedding capabilities.

4.3. Model Training

For speech encoder, we select official checkpoints of HuBERT [7] base/large/large ft, and WavLM [8] base+ /large. For each types of speech encoder, we use two types of training data to train the dialogue language model: the Fisher dataset [17] and the Fisher dataset combined with pseudo-stereo data. This approach allows us to investigate the impact of incorporating pseudo-stereo data on the performance of the spoken dialogue modeling.

We follow the same architecture and training procedure of dialogue language model and vocoder in dGSLM [13] but with different dataset and speech encoders. We selected 100 hours of training data to clustered the last layer of representations from speech encoder to encode speech into 500 clusters.

For vocoder, we utilized single-channel data from Fisher[7] combined with VCTK [29] for better generalizability of vocoder across Fisher and Podcast data.

4.4. Evaluation Metrics

4.4.1. Turn-taking Event Statistics

We follows the definition of turn-taking events in [13]. Inter-Pausal Unit (IPU) indicates the frames that a speaker is speaking; Gap is the silence frames occur between two speakers; Pause is the silence frames occur within the same speaker; Overlap indicates the frames that two speakers are both speaking.

We generated samples by promting the dialogue language model by 30 seconds dialogue from test set, and hypothesize 60 seconds continuation. The decoding temperature is set to 1.0, and sampling among the top 20 possible units. We find silence frames by applying VAD[5] to each channel separately, and compute IPU, Gap, Pause, and Overlap.

4.4.2. Human Evaluation

We conducted human evaluation to assess semantic coherence of our dialogue generation models.

We chose 30 10-second prompts each from the test sets of the Fisher dataset and the pseudo-stereo data. These prompts were carefully selected for high quality, considering factors like audio clarity and minimal noise. Then using the same sampling configuration as described in Section 4.4.1, we generated 20-second continuations for each prompt. Workers are asked to rate the samples based solely on the content and semantic coherence from 1 to 5.

We employed MTurk[8] to recruit workers. Every hit is composed by 20 samples, and each sample is evaluated by 3 workers. We rejected all workers that their work time are less than 10 minutes, or gives 1 or 2 for ground truth.

[3] We are of the view that the investigation of speech encoders can inclusively incorporate those trained on labeled data, provided they can effectively model the nuances of non-verbal speech information.

[4] https://www.apple.com/apple-podcasts/

[5] https://github.com/pyannote/pyannote-audio

[6] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[7] Generated by the script: https://gitlab.nrp-nautilus.io/ar-noc/nemo/-/blob/master/scripts/process_fisher_data.py

[8] https://www.mturk.com/

Table 1: *This table shows the turn-taking metrics. We report their difference with ground truth. Dur. indicates duration of events (seconds), and Occur. indicates occurence (times)*

	Encoder	Type	Δ IPU		Δ Gap		Δ Overlap		Δ Pause	
			Dur.	Occur.	Dur.	Occur.	Dur.	Occur.	Dur.	Occur.
(a)	Ground Truth		0 (56.86)	0 (19.86)	0 (2.61)	0 (2.88)	0 (4.29)	0 (3.96)	0 (4.83)	0 (7.42)
FISHER / FISHER + PSEUDO-STEREO ADUIO										
(b)	WavLM	base+	3.17 / -3.96	3.39 / -1.96	-0.37 / 1.02	1.49 / -0.57	2.23 / -0.83	4.47 / 0.95	-0.57 / 2.1	0.74 / 1.92
(c)		large	-3.72 / 5.36	9.96 / 2.93	0.43 / -1.00	3.58 / 0.91	0.22 / 3.14	3.72 / 4.47	3.5 / -1.24	5.54 / 0.62
(d)	HuBERT	base	4.35 / 9.38	5.66 / 5.5	-0.35 / -0.94	2.06 / 0.91	2.64 / 6.25	6.54 / 8.49	-1.26 / -2.2	-0.6 / -0.98
(e)		large	16.13 / 12.59	8.62 / 5.89	-0.92 / -1.33	1.01 / -0.28	13.94 / 10.97	11.08 / 8.62	-1.28 / -0.32	-0.81 / 0.7
(f)		large ft	1.91 / 2.56	12.2 / 10.98	0.69 / -0.07	4.79 / 3.14	2.65 / 2.57	7.24 / 7.1	0.04 / 0.07	2.68 / 3.26

Table 2: *The human evaluation of meaningfulness, with 95% confidence intervals. Semantic coherence generally improves with the addition of stereo audio, as indicated by bold scores.*

	Encoder	Type	M-MOS	
			Fisher	Podcast
(a)	Ground Truth		3.55 ± 0.23	4.2 ± 0.2
FISHER				
(b)	WavLM	base+	3.14 ± 0.23	3.31 ± 0.24
(c)		large	3.45 ± 0.25	3.17 ± 0.29
(d)	HuBERT	base	3.29 ± 0.27	3.36 ± 0.31
(e)		large ft	3.50 ± 0.31	3.67 ± 0.30
FISHER + PSEUDO-STEREO ADUIO				
(f)	WavLM	base+	**3.25 ± 0.24**	3.18 ± 0.27
(g)		large	3.33 ± 0.37	**3.49 ± 0.24**
(h)	HuBERT	base	**3.47 ± 0.25**	**3.43 ± 0.17**
(i)		large ft	**3.65 ± 0.25**	**3.74 ± 0.17**

5. Results

5.1. Turn-taking Behaviors

Table 1 presents the turn-taking events discussed in Section 4.4.1 for different settings. We report their differences from the ground truth, with a slash separating whether the pseudo-stereo data was used during training.

Our initial observations reveal that most cases (row (b), (c), (d), (f)) successfully capture turn-taking behaviours similar to real dialogue. Combining with pseudo-stereo data does not further improve the performance, which indicate that limited dialogue data are suficient to model these turn-taking behavior.

However, HuBERT large (row (e)) fail to correctly model spoken dialogue. This discrepancy primarily stems from the vocoder trained on units encoded by HuBERT large, which struggles to accurately resynthesize audio (even prompt can not be correctly reconstruct, see our demo page[2]). It suggest that the representations in the last layer of HuBERT large contains little information about original audio. Conversely, fine-tuning the model with ASR gives better phoneme discrimination for representations in the last layer, making it easier to reconstruct the original speech from the discrete units. This makes the ASR fine-tuned model more suitable for our case

We do not report the performance of HuBERT large in the following results, for it is too corrupted.

5.2. Human Evaluations

Table 2 reports the M-MOS of generated speech from prompts in Fisher and Podcast respectively. Combining pseudo-stereo data during training improves the semantic coherence within speech as seen in row (h), (i), (f) for Fisher, (g) for Podcast.

Additionally, HuBERT large ft model combined with pseudo-stereo data performs exceptionally well among prompts from both Fisher and Podcast. In fact, the performances even surpasses the ground truth of Fisher dataset.

This improvement may stem from the nature of the datasets. Fisher, being a telephonic dialogue dataset, tends to have more casual conversations with less information density. On the other hand, podcasts often have a central theme, resulting in more semantically coherent dialogue. The large scale of the pseudo-stereo data, when compared to Fisher, might also influence the models to generate dialogue resembling podcasts, thus enhancing its meaningfulness.

It's worth noting that we did not directly report the naturalness of generated audio in Table 2, although we conducted an Naturalness MOS (N-MOS) survey. We evaluate the same data as M-MOS, along with the resynthesized ground truth. The maximum difference between model generated continuation and resynthesized ground truth was no greater than 0.4; while ground truth resynthesized by different vocoder can vary by 0.8,

This observation suggests that our current resynthesis approach is not robust enough and often generates noise or distorted speech, leading to testers assessing mainly the vocoder's performance rather than the naturalness of turn-taking behavior.

6. Conclusion

This work presents a novel solution to address the scarcity of stereo dialogue data in Spoken Dialogue Modeling. Our developed pipeline transforms single-channel dialogue into pseudo-stereo data, significantly expanding our training dataset from 2,000 to 17,600 hours, and we open source our dataset for future research.

Additionally, we explored the effectiveness of using discrete units from various speech foundation models for dialogue generation. Our experiments revealed that employing HuBERT large ft as our speech encoder, and combined with pseudo-stereo data as training data notably improved the semantic coherence of the generated dialogue.

In conclusion, our contributions provide practical solutions to the challenges associated with limited stereo dialogue data and speech encoder selection. The integration of pseudo-stereo data and HuBERT large ft encoder has led to substantial improvements in spoken dialogue synthesis. However, it's impor-

tant to note that the current unit-to-speech vocoder is not yet robust enough, impacting the quality of the generated audio. This suggests a potential avenue for future research to further enhance the audio quality of speech resynthesis from discrete units.

7. References

[1] V. Nguyen, O. Versyp, C. Cox, and R. Fusaroli, "A systematic review and bayesian meta-analysis of the development of turn taking in adult–child vocal interactions," *Child Development*, vol. 93, no. 4, pp. 1181–1200, 2022.

[2] E. A. Schegloff, "Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences," *Analyzing discourse: Text and talk*, vol. 71, pp. 71–93, 1982.

[3] V. H. Yngve, "On getting a word in edgewise," in *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, 1970, pp. 567–578.

[4] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon *et al.*, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009.

[5] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu *et al.*, "Audiogpt: Understanding and generating speech, music, sound, and talking head," *arXiv preprint arXiv:2304.12995*, 2023.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "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.

[9] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[10] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[11] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhotia, T.-A. Nguyen, M. Rivière, A. Mohamed, E. Dupoux *et al.*, "Text-free prosody-aware generative spoken language modeling," *arXiv preprint arXiv:2109.03264*, 2021.

[12] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "Audiolm: a language modeling approach to audio generation," 2023.

[13] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed *et al.*, "Generative spoken dialogue language modeling," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.

[14] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.

[15] J. Sakuma, S. Fujie, and T. Kobayashi, "Response timing estimation for spoken dialog systems based on syntactic completeness prediction," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 369–374.

[16] K. Inoue, B. Jiang, E. Ekstedt, T. Kawahara, and G. Skantze, "Real-time and continuous turn-taking prediction using voice activity projection," 2024.

[17] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: A resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.

[18] H. Bredin, "pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *24th INTERSPEECH Conference (INTERSPEECH 2023)*. ISCA, 2023, pp. 1983–1987.

[19] T. J. Park, N. R. Koluguri, J. Balam, and B. Ginsburg, "Multi-scale speaker diarization with dynamic scale weighting," *arXiv preprint arXiv:2203.15974*, 2022.

[20] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[21] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[22] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, p. 1256–1266, Aug 2019. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2019.2915167

[23] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[24] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[25] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8102–8106.

[26] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[27] A. Clifton, A. Pappu, S. Reddy, Y. Yu, J. Karlgren, B. Carterette, and R. Jones, "The spotify podcast dataset," *arXiv preprint arXiv:2004.04270*, 2020.

[28] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[29] K. M. J. Yamagishi, C. Veaux, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.