Interspeech 2024
1-5 September 2024, Kos, Greece

SaSLaW: Dialogue Speech Corpus with Audio-visual Egocentric Information Toward Environment-adaptive Dialogue Speech Synthesis

Osamu Take[1], *Shinnosuke Takamichi*[1,2], *Kentaro Seki*[1], *Yoshiaki Bando*[3], *Hiroshi Saruwatari*[1]

[1]The University of Tokyo, Japan, [2]Keio University, Japan
[3]National Institute of Advanced Industrial Science and Technology (AIST), Japan
flymoons0325@g.ecc.u-tokyo.ac.jp

Abstract

This paper presents *SaSLaW*, a spontaneous dialogue speech corpus containing synchronous recordings of what speakers speak, listen to, and watch. Humans consider the diverse environmental factors and then control the features of their utterances in face-to-face voice communications. Spoken dialogue systems capable of this adaptation to these *audio environments* enable natural and seamless communications. SaSLaW was developed to model human-speech adjustment for audio environments via first-person audio-visual perceptions in spontaneous dialogues. We propose the construction methodology of SaSLaW and display the analysis result of the corpus. We additionally conducted an experiment to develop text-to-speech models using SaSLaW and evaluate their performance of adaptations to audio environments. The results indicate that models incorporating hearing-audio data output more plausible speech tailored to diverse audio environments than the vanilla text-to-speech model.

Index Terms: speech corpus, spoken dialogue, speech chain, Lombard effect, entrainment

1. Introduction

Text-to-speech (TTS) is an important technology for spoken dialogue systems (SDSs) such as conversational robots [1]. These systems are often implemented in conversational scenarios within real-world environments. Human-to-human voice communication in real environments often involves natural and intelligible speech tailored to surrounding factors such as background noise and their physical proximity. We call these environmental factors collectively as the *audio environment*.

The adaptation to audio environments by humans is based on the auditory and visual information they perceive [2, 3], which can be explained within the framework of the speech chain [4]. Reports have also indicated that different audio environments necessitate natural speech variations of conversational robots for humans [5]. Therefore, TTS incorporating audio environment inputs with the framework of speech chain is necessary for SDSs to achieve natural and seamless speech communication in dialogues. We refer to this dialogue TTS as environment-adaptive TTS (EA-TTS). Figure 1 illustrates the application of EA-TTS.

Deep neural networks (DNNs) and prevailing large-scale corpora [6] enable TTS models to generate natural speech comparable to humans for read speech in quiet environments. However, an EA-TTS model cannot be constructed only with the speaker's clean speech recorded in quiet backgrounds. EA-TTS should require first-person recordings of what humans speak, hear, and see during dialogues in various audio environments. Corpora for such EA-TTS and their construction methods are yet to be established despite the prevalence of TTS corpora.

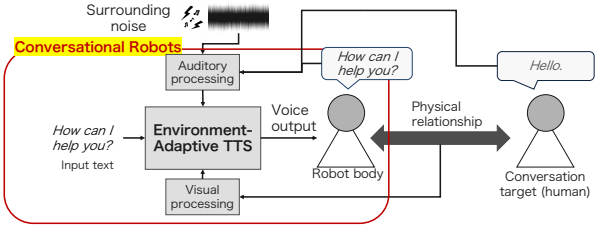

Figure 1: *Dialogue agent that generates speech using environment-adaptive TTS (EA-TTS). EA-TTS changes the speech style affected by interlocutor or environmental noise, just like us humans. This paper proposes a methodology of corpus construction to realize this and build open-source corpus.*

We present a methodology to construct a spontaneous dialogue speech corpus containing synchronous recordings of egocentric audio-visual perceptions. Following this methodology, a novel speech corpus called *SaSLaW*[1] is constructed and published[2]. We also construct EA-TTS models based on DNNs using SaSLaW and conduct a comparative evaluation.

2. Related Work

2.1. Emulating Adaptations to Audio Environments

Noise environment. The Lombard effect [7] describes the involuntary voice raising of humans in noisy environments. Several methods have been proposed for mimicking the Lombard effect, including using signal-processing-inspired manipulations [8, 9] and neural TTS [10, 11, 12]. However, previous research [12] confirmed the degradation in the naturalness of manipulated speech by signal processing. These methods are also limited to the study of read speech and do not model the Lombard effect in spontaneous dialogue [13].

Toward listeners. Prosody of human speech is also affected by the physical proximity between talkers [14] and interlocutors' speech [15], a phenomenon known as *entrainment* [16]. Several methods construct neural TTS models with architectures for adapting to interlocutors' speech [17] and faces [18]. However, none of these methods address both spontaneous dialogue scenes and the use of egocentric perceptions.

2.2. Corpora for Environment-adaptive TTS

Several corpora [19, 20] have been constructed for TTS with modeling adaptations to noisy surroundings. These corpora primarily focus on modeling read-style Lombard speech. The construction of TTS corpora modeling adaptation to environmental noises in spontaneous conversations remains unexplored.

[1]"**S**o, what **a**re you **S**peaking, **L**istening, **a**nd **W**atching?"
[2]https://github.com/sarulab-speech/SaSLaW

10.21437/Interspeech.2024-1554

Table 1: *Corpora comparison. "Spon." and "perf." are spontaneous and performative styles, respectively. "fp" and "tp" are first- and third-persons views, respectively. "IR" indicates impulse responses between talkers, and † means near-real noise.*

corpus	style	noise	hear	see	speak	IR
TTS corpus						
SaSLaW(ours)	spon.	real†	fp	fp	✓	✓
Hurricane [19]	read	real†	-	-	✓	-
CEJC [22]	spon.	real	tp	tp	✓	-
Guo et al. [21]	perf.	-	-	-	✓	-
Datasets not focusing on TTS						
EgoCom [24]	spon.	real	fp	fp	-	-
EasyCom [23]	spon.	real	tp	fp	✓	-
Hurricane 2.0 [25]	read	real†	tp	-	✓	✓

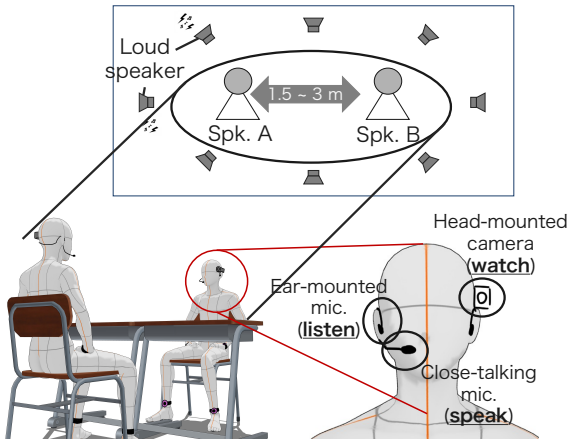

Figure 2: *Recording configuration during two-person conversation. Top illustrates the configuration of the noisy environment, and bottom illustrates participants' equipment.*

There are TTS corpora for modeling dialogue scenes used with DNN-based TTS models [21, 22]. The CEJC corpus [22] contains spontaneous conversations in real environments, while capturing environmental noise and visual footage from a third-person perspective. The EasyCom [23] dataset contains participants' speech and egocentric videos during conversations in noisy environments but lacks the variety of audio environments. EgoCom [24] contains firsthand audio-visual experiences but lacks speech recordings with minimal external noise as it was primarily designed for recognition and understanding tasks. Table 1 compares these corpora to our SaSLaW.

3. Corpus-construction Methodology

We describe the construction methodology of the spontaneous dialogue corpus SaSLaW, which includes first-person-perspective multi-modal information.

3.1. Overview of SaSLaW

SaSLaW captures adaptation to audio environments (noise, interlocutor) in spontaneous human speech communication. Achieving his involves recording scenes in which two participants engage in spontaneous dialogue while facing each other in a simulated noisy environment, mimicking real-world conditions. Figure 2 illustrates the recording configuration during a conversation between two participants. SaSLaW records what a participant speaks, listens to, and watches synchronously across two participants. Speech recordings can be utilized for constructing TTS models to generate natural speech incorporating human-like auditory and visual information.

3.2. Recording Configuration and Procedure

Participant equipment. Two participants engage in conversation in a single indoor space, referred to as the recording room. They sat facing each other across a table. The distance between the two participants is set within 1.5 to 3 m and is not strictly controlled. Participants are equipped with a close-talking microphone[3], ear-mounted binaural microphone[4], and head-mounted camera[5] on their heads, as illustrated in Figure 2. Each device corresponds to recording what they speak, hear, and see. The two microphones record at a sampling frequency of 44.1 kHz, while the head-mounted camera records at a frame rate of 30 fps. All six sensors record information synchronously. During the recording, variations in speech volume due to the audio environment are expected. To capture these variations, the gain for each participant's microphones is fixed throughout all recordings.

Environmental noise. To simulate various real environments with noise, eight loudspeakers are positioned as shown in Figure 2, covering the area around the participants. Each loudspeaker plays a different segment of the same environmental noise, simulating diffusive environmental noise in the real environment. The real-environmental-noise data are derived from a subset of the DEMAND dataset [26]. The type and power of the noise played from loudspeakers are altered after a certain number of conversation recordings. Before recording the conversation, the ambient noise level in dB is measured at the center of the two participants using a noise meter[6].

Conversation content. The two participants are instructed on the theme and roles (e.g. *sightseeing, the guide and tourist*) and tasks. During the recording, the two participants engage in improvisational conversation consisting of five to eight turns, following provided instructions.

Annotation as TTS corpus. To use the SaSLaW for TTS use, we automatically segment close-talking microphone voices using pyannote.audio [27] into utterances and transcribe them into texts using whisper [28], followed by manual correction.

3.3. Data Collection for Reproducible Evaluation

In subjective evaluations, evaluators should assess the plausibility within audio environments based on what they would listen to at the listener position while models output speech via sonic transmission, rather than synthetic speech itself. Therefore, we collect supplementary data following the previous evaluation methodology [25]. SaSLaW records impulse responses from talker to listener, positioned as shown in Figure 2, and ambient noise-only audio in the listener's position using the ear-mounted binaural microphone. Impulse responses are recorded for various distances between participants. Synthetic utterance samples are convolved with a certain impulse response and added with recorded noise-only audio. Then we acquire the evaluation samples simulating what listeners would hear.

4. Corpus Analysis

Four pairs of Japanese participants[7] engaged in the recordings of spontaneous dialogues, as outlined in Section 3. This section reports on the analysis of two pairs' SaSLaW recordings.

[3] https://www.shure.com/en-US/products/microphones/pga31
[4] https://soundprofessionals.com/product/MS-EHB-2/
[5] https://ordro.online/en-jp/products/camcamcorder-ep8
[6] https://www.sanwa.co.jp/product/syohin?code=CHE-SD1
[7] There are three male-male pairs and one female-female pair.

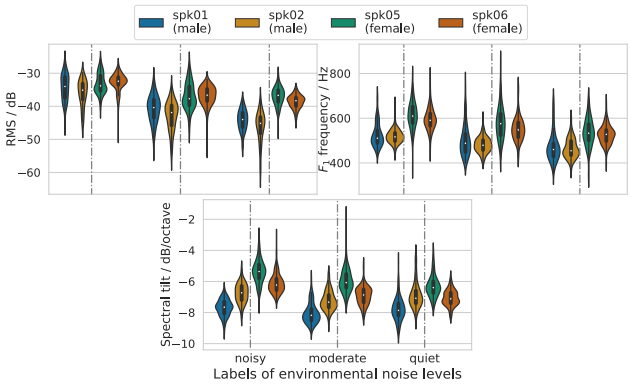

Figure 3: *Feature distributions of each speaker.*

Table 2: *The correlation coefficients of the RMS and F_0 between the speaker's utterance and the interlocutor's last utterance before the speaker's. Significant correlations are displayed* ***bold****.*

	spk01-spk02			*spk06-spk05*		
	noisy	moderate	quiet	noisy	moderate	quiet
RMS	**0.68**	**0.65**	0.07	**0.24**	**0.23**	0.00
F_0	**0.47**	0.10	0.21	**0.23**	**0.35**	0.15

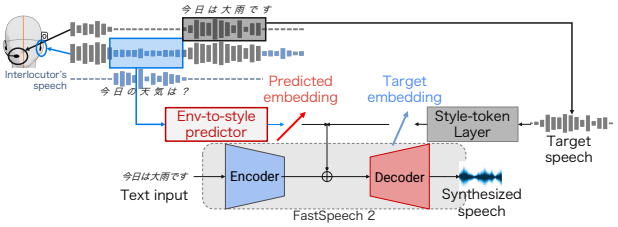

Figure 4: *Diagram of EA-TTS model. This model predicts the style vector extracted from the target utterance in training.*

One pair consisted of two male participants (named as *spk01*, *spk02*), while the other pair consisted of two female (named as *spk05*, *spk06*). The total utterance length of recorded speech was approximately 30 minutes on average.

4.1. Purpose and Procedures of Corpus Analysis

This analysis examined how spontaneous speech changed depending on the sound pressure level of environmental noise and inter-speaker coordination. Note that we did not investigate the distance between participants.

First, each conversation was assigned a label about environmental noise levels (*env-label*) such as "noisy" (high environmental noise level), "moderate" (moderate), or "quiet" (low). The label annotation was based on the sound pressure level of environmental noise measured in the conversation recording. Next, the root mean square of amplitudes (RMS), F_0, F_1 (first formant) frequency, and spectral tilt [29] were calculated for voiced frames of utterances as prosodic features. F_0 was computed and voiced frames were detected using harvest [30]. The F_1 frequency was computed through Praat [31]. Spectral tilt was computed following previous research [29], with a filter bank from 0.25 to 8 kHz. All the statistical tests below were conducted at a significance level of $p = 0.05$.

4.2. Analysis Results and Discussion

Noise. Figure 3 illustrates the distributions of RMS, F_1 frequencies and spectral tilts for each speaker. It is shown that the average F_1 frequency significantly increased from "quiet" to "noisy" env-labels for all the speakers. The average of RMS showed a significant rise from "quiet" to "noisy," for all the speakers except for spk05. The average spectral tilt presented a significant increase for female speakers from "quiet" to "noisy," whereas a different trend was observed for male.

These increases corresponded to the result reported in the previous study [20]. Also, the result suggests that while some features share characteristics among speakers in adapting to varying levels of environmental noise, others do not exhibit such commonality. It indicates that solely relying on rule-based methods with signal processing makes it difficult to accurately simulate Lombard speech adapting to audio environments.

Interlocutors. Table 2 shows the inter-speaker correlation coefficients of utterance RMS and F_0. The result indicates that in moderate and noisy environments, the target utterance features are significantly correlated with those of the last interlocutor's utterances. Adaptations to harsh listening environments may evoke this correlation enhancement from quiet to moderate, noisy. This discussion is akin to the report about the relationship between conversation excitement and entrainment in a previous study [15]. These results and discussion also suggest that SaSLaW is suitable for modeling the comprehensive consideration of and adaptation to audio environments, including noise and interlocutor factors.

5. Environment-adaptive TTS Experiment

We utilized spk01 and spk06 data to construct single-speaker EA-TTS models for each speaker. We evaluated the plausibility of synthetic speech in various audio environments.

5.1. Experimental Conditions

We compared three neural EA-TTS models. All of the models basically employed open-sourced FastSpeech 2 [32] and HiFi–GAN [33][8]. The details of these models are as follows.

- **FS2**: basic FastSpeech 2 fine-tuned on SaSLaW speech data.
- **FS2-predsty**: EA-TTS model fine-tuned on SaSLaW speech itself and auditory input during the interlocutor's last turn preceding that speech (mixed with background noises).
- **FS2-predsty-ptrn**: FS2-predsty, further pre-trained on pseudo-environment-adaptive data and fine-tuned on SaSLaW speech and hearing-audio data.

FS2 and {FS2-predsty, FS2-predsty-ptrn} models have 35M and 61M parameters, respectively. FS2 and FS2-predsty were pre-trained on the JSUT [34], an existing TTS corpus. Figure 4 illustrates the EA-TTS model applied to FS2-predsty and FS2-predsty-ptrn. EA-TTS incorporates a style-token layer and Env-to-style predictor into the original FastSpeech 2. The style-token layer uses global style token [35], which extracts a style vector with a fixed length from an utterance. The Env-to-style predictor consists of four trainable convolution layers and an energy extractor, which predict the style vector from hearing audio. The objective L is defined as $L = L_{\mathrm{TTS}} + L_{\mathrm{sty}}$, where L_{TTS} denotes the objective of FastSpeech 2 and L_{sty} denotes the L1 loss between style vectors.

Pseudo-data training. We created a pseudo-environment-adaptive dataset by using a TTS corpus and noise dataset

[8] https://github.com/Wataru-Nakata/FastSpeech2-JSUT. We also followed its hyperparameter settings.

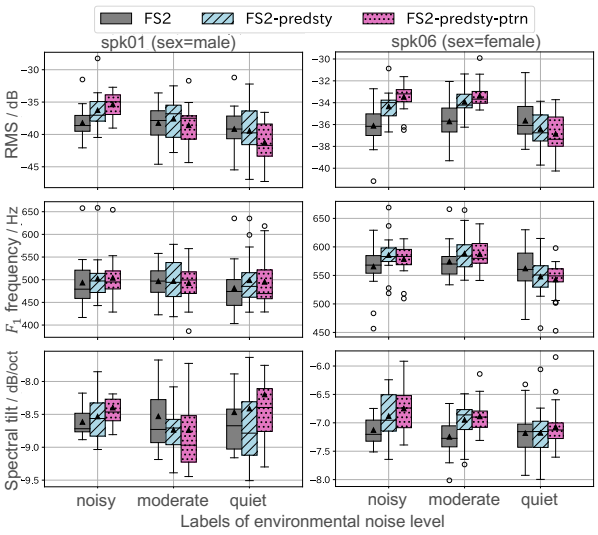

Figure 5: *Feature distributions of synthetic speech. Solid triangles indicate the average of distributions, and black lines indicate centroid.*

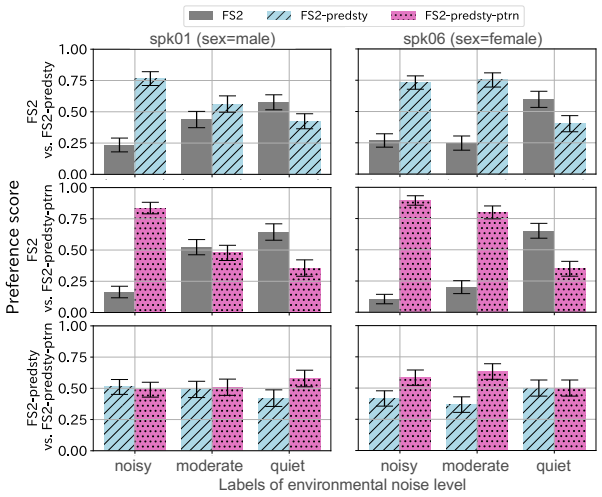

Figure 6: *Preference score of each model pair. Error bars denote 95% confidence intervals.*

through signal processing. While utterances manipulated in a signal-processing manner deviate in features from real Lombard speech, these utterances can effectively pretrain EA-TTS models due to their scalability. We assign each utterance an arbitrary noise signal at a varying level then increase its spectral tilt to enhance intelligibility within the corresponding noise. We use JSUT as the TTS corpus and DEMAND as the noise dataset.

Training conditions. We split SaSLaW audio recordings into train and test sets for each speaker's data. There was no overlap in the environmental noise contained in the hearing audio between the sets. The test set covered all env-labels assigned to speech. Finally, we split the recordings of spk01 into 299/49 and spk06 into 443/64 train/test utterances. All three EA-TTS models were trained on a single NVIDIA GeForce RTX 4090 GPU. They were pre-trained for 900k ($\leq$ three days) and fine-tuned on single-speaker recordings of SaSLaW for 100k steps ($\leq$ 12 hours).

5.2. Objective Evaluation

For objective evaluation, we computed the prosodic features outlined in Section 4.1 for the synthetic speech. We investigated whether the distribution changed across different env-labels, consistent with the analysis presented in Figure 3.

Figure 5 illustrates the prosodic feature distributions of synthetic speech. FS2 showed no significant differences across env-labels. For FS2-predsty and FS2-predsty-ptrn, the synthetic speech of spk01 showed a significant increase only for RMS from "quiet" to "noisy". For spk06, the synthesized speech exhibited a significant increase from "quiet" to "noisy" for all features except between "moderate" and "noisy" for spectral tilt. These results indicate that the Env-to-style predictor enabled the generation of speech with characteristics adapted to audio environments.

5.3. Subjective Evaluation

We conducted an AB preference test to compare the plausibility of the evaluation speech samples within surrounding noises. Each synthetic utterance was processed as described in Section 3.3 to serve as the evaluation utterance. Evaluators selected which utterance sounded more plausible in environmental noises through the listening experiment. The evaluators were provided with a combined criterion of naturalness and intelligibility as supplementary instruction. For each model-pair configuration, 72 evaluators were recruited and 720 responses were collected via Lancers[9]. Figure 6 illustrates the env-label-wise preference scores for each model-pair and speaker.

The results indicate that for the "noisy" label, both FS2-predsty and FS2-predsty-ptrn significantly outperformed FS2 in preference scores for both speakers, while FS2 significantly surpassed the other two for the "quiet" label. This result suggests that the EA-TTS models with the Env-to-style predictor successfully adapted to noisy surroundings. However, this adaptation to "quiet" environments degraded the plausibility of synthetic speech compared with FS2-synthesized speech, which had averaged prosodic features and gained intelligibility.

Figure 6 shows that the preference scores of FS2-predsty-ptrn were equal to or significantly higher than those of FS2-predsty. This suggests that pseudo-data pre-training improved the performance.

6. Conclusion

We introduced SaSLaW, a novel speech corpus for generative tasks with synchronized audio and visual first-person recordings. We described the methodology of constructing SaSLaW and analyzed the recordings to confirm human speech's adaptations to audio environments. The experimental results indicate that SaSLaW enables the construction of environment-adaptive TTS models by using the auditory perception of the target speaker as input, successfully producing plausible speech tailored to diverse audio environments. This work does not explore the analysis and modeling of speech adaptation to visual information, which can be investigated for further work.

7. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 23H03418 and 22H03639, Moonshot R&D Grant

[9] https://www.lancers.jp

Number JPMJPS2011, and JST FOREST JPMJFR226V. The authors also thank Aya Watanabe for her support in designing the figures for this paper.

8. References

[1] T. Kawahara, "Spoken dialogue system for a human-like conversational robot ERICA," in *Proc. International Workshop on Spoken Dialogue System Technology (IWSDS)*, 2019, pp. 65–75.

[2] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, 2014.

[3] V. Hazan and R. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions." *The Journal of the Acoustical Society of America*, vol. 130 4, pp. 2139–52, 2011.

[4] P. B. Denes and E. Pinson, *The speech chain*. Macmillan, 1993.

[5] P. Tuttosi, E. Hughson, A. Matsufuji, C. Zhang, and A. Lim, "Read the room: Adapting a robot's voice to ambient and social contexts," in *Proc. 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3998–4005.

[6] X. Tan, *Neural Text-to-Speech Synthesis*. Springer Nature, 2023.

[7] E. Lombard, "Le signe de l'élévation de la voix," *Annales des Maladies de L'Oreille et du larynx*, vol. 37, pp. 101–119, 1911.

[8] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. INTERSPEECH*, 2012, pp. 635–638.

[9] T. V. Ngo, R. Kubo, and M. Akagi, "Mimicking Lombard effect: An analysis and reconstruction," *IEICE Transactions on Information and Systems*, vol. E103.D, no. 5, pp. 1108–1117, 2020.

[10] B. Bollepalli, L. Juvela, and P. Alku, "Lombard speech synthesis using transfer learning in a Tacotron text-to-speech system," in *Proc. INTERSPEECH*, 2019, pp. 2833–2837.

[11] S. Novitasari, S. Sakti, and S. Nakamura, "A machine speech chain approach for dynamically adaptive Lombard TTS in static and dynamic noise environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2673–2688, 2022.

[12] T. Raitio, P. Petkov, J. Li, M. Shifas, A. Davis, and Y. Stylianou, "Vocal effort modeling in neural TTS for improving the intelligibility of synthetic speech in noise," in *Proc. INTERSPEECH*, 2022, pp. 1936–1940.

[13] F. S. Laura Folk, "The Lombard effect in spontaneous dialog speech," in *Proc. INTERSPEECH*, 2011, pp. 2701–2704.

[14] D. Pelegrín-García, B. Smits, J. Brunskog, and C.-H. Jeong, "Vocal effort with changing talker-to-listener distance in different acoustic environments," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 1981–1990, Apr. 2011.

[15] R. Nishimura, N. Kitaoka, and S. Nakagawa, "Analysis of factors to make prosodic change in spoken dialog (feature articles; rhythm and timing)," *Journal of the Phonetic Society of Japan*, vol. 13, no. 3, pp. 66–84, 2009.

[16] R. Levitan, A. Gravano, L. Willson, B. Štefan, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, Jun. 2012, pp. 11–19.

[17] J. Li, Y. Meng, X. Wu, Z. Wu, J. Jia, H. Meng, Q. Tian, Y. Wang, and Y. Wang, "Inferring speaking styles from multi-modal conversational context by multi-scale relational graph convolutional networks," in *Proc. ACM International Conference on Multimedia*, ser. MM '22, 2022, p. 5811–5820.

[18] M. Zhou, Y. Bai, W. Zhang, T. Yao, T. Zhao, and T. Mei, "Visual-aware text-to-speech*," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[19] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. INTERSPEECH*, August 2013, pp. 1341–1345.

[20] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual Lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, Jun. 2018.

[21] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational end-to-end TTS for voice agents," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 403–409.

[22] H. Koiso, H. Amatani, Y. Den, Y. Iseki, Y. Ishimoto, W. Kashino, Y. Kawabata, K. Nishikawa, Y. Tanaka, Y. Usuda, and Y. Watanabe, "Design and evaluation of the corpus of everyday Japanese conversation," in *Proc. Language Resources and Evaluation Conference (LREC)*. European Language Resources Association, Jun. 2022, pp. 5587–5594.

[23] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv preprint arXiv:2107.04174*, 2021.

[24] C. G. Northcutt, S. Zha, S. Lovegrove, and R. Newcombe, "EgoCom: A multi-person multi-modal egocentric communications dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 6783–6793, 2023.

[25] J. Rennies, H. Schepker, C. Valentini-Botinhao, and M. Cooke, "Intelligibility-enhancing speech modifications — The Hurricane Challenge 2.0," in *Proc. INTERSPEECH*, 2020, pp. 1341–1345.

[26] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments." Zenodo, Apr. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1227121

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

[28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. International Conference on Machine Learning (ICML)*, ser. ICML'23, 2023, pp. 28 492–28 518.

[29] Y. Sato and J. Villegas, "Spectral tilt may have a smaller impact on the intelligibility of speech in noise," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–5.

[30] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. INTERSPEECH*, 2017, pp. 2321–2325.

[31] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.1.38)," 2021. [Online]. Available: http://www.praat.org

[32] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[33] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. International Conference on Neural Information Processing Systems (NeurIPS)*, ser. NIPS'20, 2020, pp. 17 022–17 033.

[34] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," *arXiv preprint arXiv:1711.00354*, 2017.

[35] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 80, 2018, pp. 5167–5176.