

# Voice Conversion for Lombard Speaking Style with Implicit and Explicit Acoustic Feature Conditioning

*Dominika Woszczyk<sup>1\*</sup>, Manuel Sam Ribeiro<sup>2</sup>, Thomas Merritt<sup>2\*</sup>, Daniel Korzekwa<sup>2\*</sup>*

<sup>1</sup>Department of Computing, Imperial College London, London, UK

<sup>2</sup>Amazon Alexa, TTS Research, UK

<sup>1</sup>d.woszczyk19@imperial.ac.uk

## Abstract

Text-to-Speech (TTS) systems in Lombard speaking style can improve the overall intelligibility of speech, useful for hearing loss and noisy conditions. However, training those models requires a large amount of data and the Lombard effect is challenging to record due to speaker and noise variability and tiring recording conditions. Voice conversion (VC) has been shown to be a useful augmentation technique to train TTS systems in the absence of recorded data from the target speaker in the target speaking style. In this paper, we are concerned with Lombard speaking style transfer. Our goal is to convert speaker identity while preserving the acoustic attributes that define the Lombard speaking style. We compare voice conversion models with implicit and explicit acoustic feature conditioning. We observe that our proposed implicit conditioning strategy achieves an intelligibility gain comparable to the model conditioned on explicit acoustic features, while also preserving speaker similarity.

**Index Terms:** speech intelligibility, speaking style conversion, Lombard speech, text-to-speech (TTS).

## 1. Introduction

Traditional approaches to improving speech intelligibility for individuals with hearing aids primarily focus on signal processing and amplification at the receiving end. However, in this paper, we propose an alternative approach that seeks to enhance intelligibility by addressing the source of speech generation rather than relying solely on signal processing and amplification. Specifically, we investigate the use of Lombard speech in text-to-speech (TTS) systems to mimic the natural adjustments made by speakers in noisy environments. We argue Lombard-style TTS could be an effective way to improve the intelligibility of speech, easier to receive with hearing loss and also more generally speech in noisy environments.

In fact, the Lombard speaking style has already been applied to Text-to-Speech (TTS) systems [1, 2, 3] and has demonstrated a positive impact on the intelligibility of synthesized voices in noise [3]. Nevertheless, current TTS models require data for the target speaker in the target speaking style, and Lombard speech is difficult to record. The recording conditions are tiring due to the noise, and the Lombard effect patterns and intensity vary from one speaker to another.

In this paper, we present a data augmentation technique for Lombard text-to-speech systems. Voice conversion has been shown to be useful as a data augmentation technique for TTS, style conversion and source style transfer [4, 5, 6] However, most voice conversion models focus on the speaker identity and ignore additional characteristics that are transferred with it (emphasis,

emotion, intonation). Voice conversion for speaking styles is challenging as it often loses the source style. We focus on the inherent challenges of performing voice conversion for Lombard speech, in the optic to train a TTS model on it and address the problem of voice conversion preserving source Lombard speaking style using implicit and explicit acoustic feature conditioning. Our goal is to convert speaker identity while retaining the source speakers' Lombard speaking style. By preserving the Lombard features of the original speaker in voice conversion, we can generate synthetic datasets for a target voice and ensure that the resulting audio signal sounds more intelligible and natural to the listener. Past works perform source style transfer by explicitly modeling key characteristics of the source style to preserve them [7, 8, 9, 10, 11]. However, this approach requires manual analysis and under-utilize the feature extraction capabilities of today's deep neural networks. More recently, models that implicitly learn the desired target style speech attributes were explored, either via adversarial loss [12, 13] or disentanglement [14, 15] for both VC and TTS. Style reconstruction loss is another technique that also has been applied to enforce expressive speech synthesis on TTS [16]. In the domain of Lombard-style voice conversion, implicit feature modeling has yet to be explored.

In our work, we focus on the problem of intelligibility-preserving voice conversion for Lombard speaking style transfer and we propose to model the Lombard prosody implicitly using a style reconstruction loss to overcome these issues and compare it to explicit modeling. We compare implicit and explicit conditioning on a many-to-many voice conversion model. Our results show that a model with a style reconstruction loss achieves an intelligibility increase comparable to a model with explicit conditioning, whilst better preserving speaker identity.

## 2. Analysis of the Lombard Speaking Style

### 2.1. Background

Signal processing approaches that are noise independent such as spectral shaping (SS) and dynamic range compression (DRC) have shown to be helpful in improving speech intelligibility [17]. SS distributes the energy in the frequency domain, sharpens the formants and reduces the spectral tilt while DRC amplifies quiet sounds and reduces loud sounds. While spectral shaping and dynamic range compression (SSDRC) can efficiently improve intelligibility without additional data collection, they do impact the naturalness of the speech samples. Lombard speech has been shown to improve the intelligibility over natural speech [18] and has been studied in the context of improving intelligibility of speech disorders and hearing aids [19], or build robust automatic speech recognition systems [18]. The Lombard effect has a high variability as it not only comes from the surrounding noise but also relies on the feedback received from the listener, and

---

\*Work done while at Amazon Alexa TTS Research.

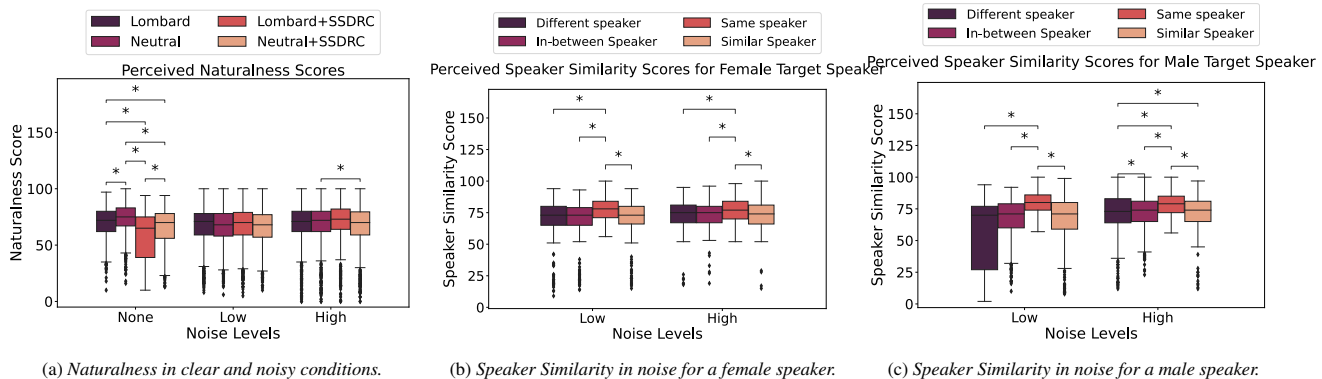


Figure 1: Boxplots of Naturalness and Speaker Similarity in both low (SNR-1) and high (SNR-3) noise levels from the Pilot study. The symbol \* indicates significance for  $p \leq 0.05$ .

the speaker itself. However, general trends have been observed. Speakers tends to increase the duration of vowels and decrease their spectral tilt; increase fundamental frequency ( $f_0$ ); and the spectral energy of vowels is moved from high and low frequencies to mid-level frequencies. In the context of TTS systems, it is important to achieve pleasant voices for users. The impact of the Lombard effect on naturalness has been observed in recent work [2, 20], but there has been no study that made it its focus. Another relevant dimension to style conversion in speech synthesis is speaker similarity and voice identity in noise. Humans are very good at detecting many cues from a person’s voice and can distinguish similar speakers given similar conditions [21]. Unfortunately studies have shown that they are not reliable when it comes to grouping speaker identity when confronted with speech samples that include different emotions, whisper, laughs or other languages [22]. Different listeners also have different levels of perception and can be further affected by noise [23].

In this section we investigate the impact of the Lombard effect in perceived intelligibility and naturalness in noise, and compare it against SSDRC. Additionally, we study the ability of listeners to perceive the speaker identity and similarity in noise. To this end, we design several pilot studies and describe them in the following sections.

## 2.2. Evaluation Setup

We perform our evaluation on the Audio-Visual Lombard Grid dataset [24] which consists of 54 speakers (24 Male, 30 Female). Each speaker has a total of 100 recordings of randomly generated sentences in English (50 Lombard and 50 neutral). We evaluate the Objective Intelligibility (OI) of our systems by computing the speech intelligibility in bits (SIIB) score [25], an objective metric for intelligibility in noise, given speech-shaped noise at two speech-to-noise ratio (SNR): SNR -1 and SNR -3. We also run subjective evaluations for Perceived Intelligibility, Naturalness and Speaker Similarity with MUSHRA-like (MULTiple Stimuli with Hidden Reference and Anchor) [26] listening tests. To measure Speaker Similarity, we ask 50 listeners to compare and rate the similarity of the systems samples to a reference sample of the target speaker. For Intelligibility and Naturalness, we do not include a reference sample but ask listeners to rate how intelligible or natural the samples sound on a scale from 0 to 100 (100 = “Very Intelligible/Natural”). For each listening test, we include 10 samples from 10 speakers, balancing for gender, for a total of 100 samples.

## 2.3. Results

Systems	SNR -1		SNR -3	
	OI	SI	OI	SI
Neutral	136.30 (.99)	57.26 (1.68)*	69.55 (.67)	26.17 (2.10)†
Lombard	154.24 (.96)	65.66 (1.37)	85.22 (.73)	26.03 (2.05)*†
Neutral + SSDRC	243.76 (1.14)	57.28 (1.69)*	148.91 (.86)	25.48 (2.02)*
Lombard + SSDRC	<b>251.78 (1.21)</b>	<b>66.36 (1.31)</b>	<b>150.32 (.82)</b>	<b>42.61 (2.33)</b>

Table 1: Objective (OI) and subjective (SI) Intelligibility in noise at SNR -1 (dB) and SNR -3 (dB) for the pilot study, measured with SIIB and MUSHRA-like test. The higher the value the better. The best system is highlighted in bold. All pairs of systems are significantly different for  $p \leq 0.005$ , except the row-wise pairs indicated by \* and †.

We compare neutral and Lombard speech samples with and without SSDRC. First, we aim to investigate the **Perceived Intelligibility** of speech in noise. Results are presented in Table 1. We observe that the Lombard effect is rated more intelligible by human listeners at low noise level (SNR -1), and by the SIIB score than *Neutral* for both SNR -1 and SNR -3. On the other hand, while the SIIB score seems to favour the SSDRC method, *Lombard* is still ranked higher than *Neutral+SSDRC* by listeners at low and high noise level. Nevertheless, the combination of *Lombard+SSDRC* performs the best in terms of both subjective and objective scores for all noise levels.

We evaluate the impact of Lombard speech on **Naturalness** in a similar fashion to the Perceived Intelligibility. Results are shown on Figure 1.a. We notice that while the Lombard effect is rated as less natural, it is still better ranked than *SSDRC* approaches. However, this effect is less apparent for high noise level. These results support the hypothesis that, in noisy conditions, the importance of naturalness is lost in exchange for intelligibility [2, 20]. This indicates that while the Lombard effect might affect the naturalness, it is negligible in noise and remains advantageous for its intelligibility gains.

Finally, we aim to understand the importance of preserving the target speaker identity during style conversion in noisy conditions, by investigating how listeners perceive **Speaker Similarity** in noise. To this aim, we pick samples from one speaker and compare them to samples from three others: a speaker with a similar voice (*Same Speaker*), one with a distinctly different voice (*Different Speaker*), and speaker that lies in-between (*In-between Speaker*). Results for both genders are presented in Figures 1.b and 1.c. We observe that even in noise, the order of similarity between speakers remains the same. Unlike the **Natu-**

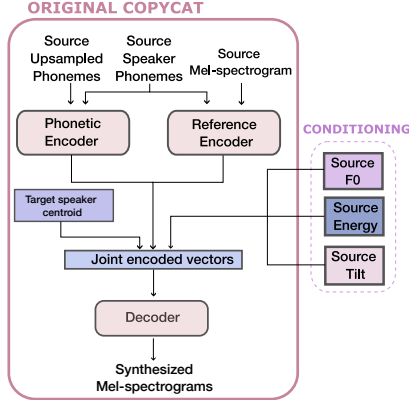


Figure 2: Overall architecture of the voice conversion model, a modified version of Copycat [4, 5] with source features conditioning.

ralness study, these results indicate that for style conversion, the speaker identity remains important to preserve.

### 3. Lombard Style Transfer

#### 3.1. Voice Conversion Model

Our voice conversion model is based on CopyCat [27], a non-parallel many-to-many prosody transfer model. As shown in Figure 2, the model has a reference encoder that takes mel-spectrograms and speaker embeddings extracted from a speaker verification model. The speaker embeddings are also passed together with phonemes upsampled at the frame level. The resulting encoded vectors are joined and given as input to a decoder that outputs decoded mel-spectrogram ready to be vocoded. Our baseline and conditioned models are the Copycat models as described in [4, 5]. We use a Kullback–Leibler (KL)-divergence loss ( $L_{KL}$ ) for the Variational Autoencoder (VAE) component in the reference encoder and an L1 loss on the source and reconstructed mel-spectrograms ( $L_{rec}$ ).

#### 3.2. Voice Conversion with Explicit Conditioning

We implement explicit acoustic conditioning for Lombard speaking style transfer, shown in Figure 2. We consider  $f_0$ ,  $mgc0$  (spectral energy) and  $mgc1$  (spectral tilt). We extract features at frame-level using WORLD vocoder [28] directly from the source speech waveform and feed them to the decoder, to condition the model to follow the source distributions. The target speaker embedding is the centroid of all embeddings from that speaker’s training data. We join them to the encoded feature vectors and feed it to the decoder.

#### 3.3. Voice Conversion with Style Reconstruction Loss

To enforce implicit feature learning, we add an auxiliary loss to our VC model, as shown in Figure 3. This loss forces the model to generate mel-spectrograms that preserve the source style, given the features that the style classifier learned representations during the pre-training. The training is split into two stages. First, a style classifier is trained on the Lombard/neutral recordings. Then, the weights of the classifier are frozen and the model is used to predict the class of synthesized mel-spectrograms. We use binary cross-entropy loss, which we call style reconstruction loss ( $L_s$ ), in addition to  $L_{rec}$  and  $L_{KL}$ . During inference, the style classifier is dropped. Finally, we evaluate a fusion model

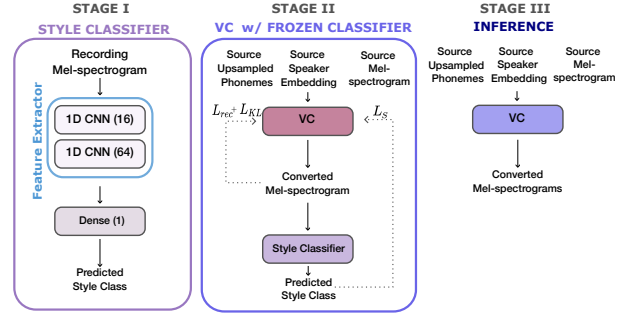


Figure 3: Architecture and training schedule of the Voice Conversion (VC) Model with style classifier. Stage I is classifier training. Stage II is training VC with classifier as frozen style discriminator, and Stage III is inference with trained VC model and the style classifier is dropped.

of style classifier and explicit conditioning by adding the source features from Section 3.2 to the model with style reconstruction loss.

#### 3.4. Evaluation Setup & Results

In our experiments we aim to answer the following questions: 1) “Can we preserve the key properties of the Lombard effect during voice conversion”; and 2) “Can we preserve speaker identity”. To evaluate our systems, we follow the same dataset and metrics setup as described in Section 2.2. We train our model on 52 speakers and pick two target speakers, one female (s27) and one male (s43) speaker with high SIIB score for their neutral recordings. We train our VC model with and w/o conditioning on a many-to-many task for 100k steps. The classifier is trained on Oracle mel-spectrograms of the Lombard Grid dataset for 5k steps. We train the VC model together with the classifier with frozen weights for an additional 100k. We use a batch size of 16 and a learning rate of 0.0001.

We summarize the results for the objective intelligibility measure for explicit and implicit conditioning systems as well as their fusion in Table 2. Figures 4 and 6 present the results of the subjective intelligibility in high noise and speaker similarity evaluation. We perform an independent t-test with a Holm-Bonferroni correction for all evaluations. We also evaluate them in low noise settings, and we observed that adding source features or enforcing the style reconstruction helped to improve the intelligibility of the samples for all systems, as all models were significantly better than the base VC model. However, in high noise settings the improvements are less apparent, as shown in Figure 4. We observe differences given the source speaker’s gender. For female source speakers, the best system is  $mgc0+mgc1$ , and  $f_0$  seems to be harmful to intelligibility. On the other hand, for male source speakers, both  $VC + L_s$  and  $f_0+mgc0+mgc1$  are significantly better than the baseline. In general,  $mgc0$  and  $mgc1$  features perform the best. However, the  $f_0$  seems to be beneficial for male source speakers. Nevertheless, looking at Figure 4 and Table 2, the model with implicit conditioning performs similarly to  $L_s+f_0+mgc0+mgc1$ . Additionally, looking at speaker similarity in Figure 6, we see that the baseline VC transfers more of the target speaker identity but that  $L_s$  perform similarly to other models. One can see on Figure 5 that combining only some of the few explicit features with the  $L_s$  loss seems to either not add much or be detrimental to the intelligibility. On the other hand, the fusion of all of them improves the performance with the best results achieved with  $L_s+mgc0+mgc1$ .

Systems	Female target		Male target	
	SNR-1	SNR-3	SNR-1	SNR-3
Source Lombard Recordings	154.25 (.96)	85.22 (.72)	134.53 (5.63)	73.18 (4.83)
VC	83.92 (3.56)*	46.29 (2.74)*	102.58(6.43)	38.78 (3.77)*
+f0	84.21 (3.56)*	46.61 (3.11)*	98.46 (5.92)	38.13 (3.10)*
+mgc0+mgc1	100.38 (1.80)	55.42 (1.47)	117.15 (5.7)*	46.91 (3.00)
+f0+mgc0+mgc1	93.85 (3.57)	53.84 (2.69)	120.73 (5.87)	49.37 (3.24)
VC + L <sub>S</sub>	86.79 (3.57)	51.44 (2.8)	107.26 (4.88)	44.14 (3.27)
+f0	106.19 (4.27)	55.23 (2.29)	<b>118.85 (7.64)*</b>	<b>51.89 (3.05)</b>
+mgc0+mgc1	104.37 (2.40)	58.24 (1.97)	87.48 (3.48)	48.83 (3.28)†
+f0+mgc0+mgc1	<b>115.13 (3.19)</b>	<b>60.65 (1.90)</b>	110.61 (6.95)	49.00 (3.72)‡

Table 2: Mean Objective Intelligibility score (SIIB) and Confidence Interval (CI) for speech-shape noise at SNR -1 and SNR -3 for female and male target speaker. Higher is better. The best results are highlighted in bold. All pairs of systems are significantly different for  $p \leq 0.005$ , except the ones indicated by \* and †.

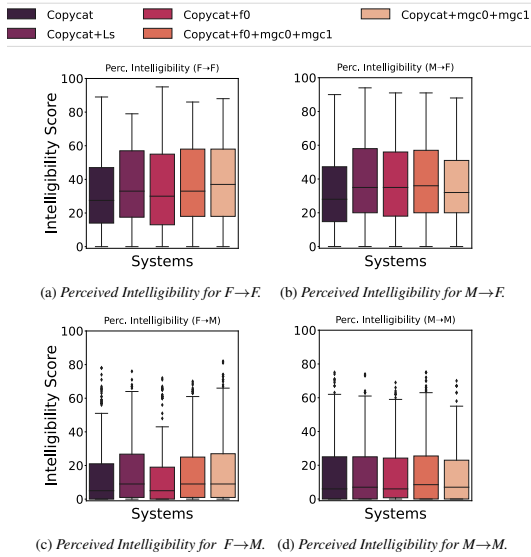


Figure 4: Cross-gender Perceived Intelligibility in high (SNR-3) noise settings of a VC model with different explicit conditioning and with style reconstruction loss ( $L_S$ ). Higher is better.

## 4. Discussion

The results show that *Implicit modeling* via the Lombard style classifier was able to achieve similar results to models with explicit conditioning. This approach presents the advantage of not requiring domain knowledge and extensive linguistic and acoustics studies. We also observe that  $f_0$  extracted from male speakers in Lombard style has better impact on the intelligibility than from female speakers, which is consistent with the gender-specific changes in Lombard speech [18]. On the other hand, our experiments on explicit modeling confirm previous studies on the importance of spectral tilt and energy for modeling the Lombard effect [29, 18]. We note that our models do not alter the durations, albeit observed in Lombard speech, which could further benefit the intelligibility. Additionally, this work is limited by the size of the dataset and the overall robotic prosody of the samples given the recording task, which also impacts the quality of the synthesized samples. Future work could explore the intelligibility enhancement on a larger and conversational recordings dataset. This in turn would allow us to train and evaluate TTS models trained on synthetic Lombard data generated with our model. We would also evaluate the intelligibility of systems on transcriptions and intelligibility scores targeted at hearing

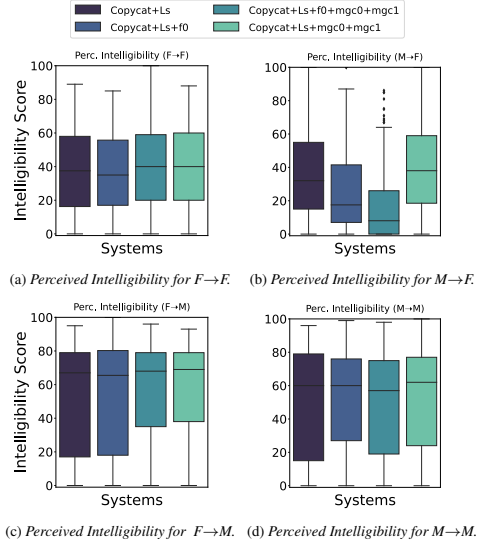


Figure 5: Cross-gender Perceived Intelligibility in high (SNR-3) noise settings of the fusion of a VC model with style reconstruction loss ( $L_S$ ) and explicit conditioning. Higher is better.

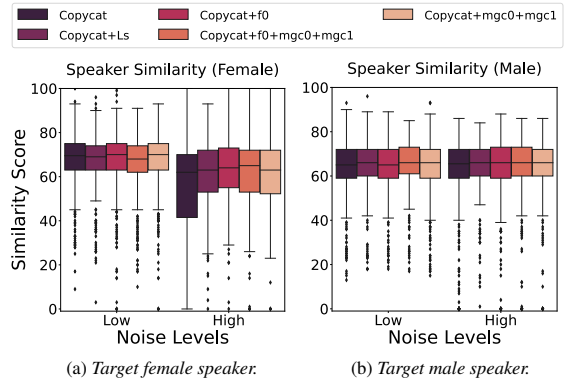


Figure 6: Perceived Speaker Similarity in low (SNR-1) and high (SNR-3) noise settings of a VC model with explicit conditioning and with style reconstruction Loss ( $L_S$ ). Higher is better.

loss. Finally, adversarial learning could also be explored in future improvements to further disentangle source speaker identity from the explicit features, as well as grouping different source speakers by Lombard intensity to control the intelligibility level.

## 5. Conclusion

In this work, we analyze the impact of the Lombard effect on the Intelligibility of voices in noise and investigate Lombard-preserving voice conversion. Confirming previous studies, we show that the Lombard effect increases the intelligibility in noise, and that while naturalness is lost, speaker similarity can still be observed by listeners in noisy conditions. We investigate many-to-many voice conversion preserving Lombard style with both implicit and explicit conditioning. Spectral tilt and energy were the most beneficial features for the Lombard style, and we show that the model with the added reconstruction loss achieves intelligibility gain on par with the model conditioned on source features, while better preserving the speaker similarity.

## 6. References

- [1] Bajibabu Bollepalli, Lauri Juvola, and Paavo Alku, “Speaking style adaptation in text-to-speech synthesis using sequence-to-sequence models with attention,” *arXiv preprint arXiv:1810.12051*, 2018.
- [2] Qiong Hu, Tobias Bleisch, Petko Petkov, Tuomo Raitio, Erik Marchi, and Varun Lakshminarasimhan, “Whispered and lombard neural speech synthesis,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 454–461.
- [3] Dipiyoti Paul, Muhammed PV Shifas, Yannis Pantazis, and Yannis Stylianou, “Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion,” *arXiv preprint arXiv:2008.05809*, 2020.
- [4] Manuel Sam Ribeiro, Julian Roth, Giulia Comini, Goeric Huybrechts, Adam Gabrys, and Jaime Lorenzo-Trueba, “Cross-speaker style transfer for text-to-speech using data augmentation,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6797–6801.
- [5] Goeric Huybrechts, Thomas Merritt, Giulia Comini, Bartek Perz, Raahil Shah, and Jaime Lorenzo-Trueba, “Low-resource expressive text-to-speech using data augmentation,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6593–6597.
- [6] Ryo Terashima, Ryuichi Yamamoto, Eunwoo Song, Yuma Shirahata, Hyun-Wook Yoon, Jae-Min Kim, and Kentaro Tachibana, “Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation,” *arXiv preprint arXiv:2204.10020*, 2022.
- [7] Gang Li, Xiaochen Wang, Ruimin Hu, Huyin Zhang, and Shanfa Ke, “Normal-to-lombard speech conversion by lstm network and bgmm for intelligibility enhancement of telephone speech,” in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.
- [8] Ana Ramírez López, Shreyas Seshadri, Lauri Juvola, Okko Räsänen, and Paavo Alku, “Speaking style conversion from normal to lombard speech using a glottal vocoder and bayesian gmms,” in *Interspeech*, 2017, pp. 1363–1367.
- [9] D-Y Huang and EP Ong, “Lombard speech model for automatic enhancement of speech intelligibility over telephone channel,” in *2010 International Conference on Audio, Language and Image Processing*. IEEE, 2010, pp. 429–434.
- [10] Tuomo Raitio, Antti Suni, Martti Vainio, and Paavo Alku, “Analysis of hmm-based lombard speech synthesis,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [11] Tuomo Raitio, Petko Petkov, Jiangchuan Li, Muhammed Shifas, Andrea Davis, and Yannis Stylianou, “Vocal effort modeling in neural tts for improving the intelligibility of synthetic speech in noise,” *arXiv preprint arXiv:2203.10637*, 2022.
- [12] Shreyas Seshadri, Lauri Juvola, Paavo Alku, Okko Räsänen, et al., “Augmented cyclegrams for continuous scale normal-to-lombard speaking style conversion,” in *Interspeech*, 2019, pp. 2838–2842.
- [13] Bastian Schnell, Goeric Huybrechts, Bartek Perz, Thomas Drugman, and Jaime Lorenzo-Trueba, “Emocat: Language-agnostic emotional voice conversion,” *arXiv preprint arXiv:2101.05695*, 2021.
- [14] Chak Ho Chan, Kaizhi Qian, Yang Zhang, and Mark Hasegawa-Johnson, “Speechsplit2. 0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6332–6336.
- [15] Yanyao Bian, Changbin Chen, Yongguo Kang, and Zhenglin Pan, “Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis,” *arXiv preprint arXiv:1904.02373*, 2019.
- [16] Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li, “Expressive tts training with frame and style reconstruction loss,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1806–1818, 2021.
- [17] Jan Rennie, Henning F Schepker, Cassia Valentini-Botinhao, and Martin Cooke, “Intelligibility-enhancing speech modifications-the hurricane challenge 2.0,” in *INTERSPEECH*, 2020, pp. 1341–1345.
- [18] Jean-Claude Junqua, “The lombard reflex and its role on human listeners and automatic speech recognizers,” *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.
- [19] John HL Hansen, Jaewook Lee, Hussnain Ali, and Juliana N Saba, “A speech perturbation strategy based on “lombard effect” for enhanced intelligibility for cochlear implant listeners,” *The Journal of the Acoustical Society of America*, vol. 147, no. 3, pp. 1418–1428, 2020.
- [20] Jennifer M Vojtech, Jacob P Noordzij Jr, Gabriel J Cler, and Cara E Stepp, “The effects of modulating fundamental frequency and speech rate on the intelligibility, communication efficiency, and perceived naturalness of synthetic speech,” *American journal of speech-language pathology*, vol. 28, no. 2S, pp. 875–886, 2019.
- [21] Marianne Latinus and Pascal Belin, “Human voice perception,” *Current Biology*, vol. 21, no. 4, pp. R143–R145, 2011.
- [22] Nadine Lavan, Luke FK Burston, Paayal Ladwa, Siobhan E Merriam, Sarah Knight, and Carolyn McGettigan, “Breaking voice identity perception: Expressive voices are more confusable for listeners,” *Quarterly Journal of Experimental Psychology*, vol. 72, no. 9, pp. 2240–2248, 2019.
- [23] Judy H Song, Erika Skoe, Karen Banai, and Nina Kraus, “Perception of speech in noise: neural correlates,” *Journal of cognitive neuroscience*, vol. 23, no. 9, pp. 2268–2279, 2011.
- [24] Najwa Alghamdi, Steve Maddock, Ricard Marxer, Jon Barker, and Guy 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, 2018.
- [25] Steven Van Kuyk, W Bastiaan Kleijn, and Richard C Hendriks, “An instrumental intelligibility metric based on information theory,” *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2017.
- [26] B Series, “Method for the subjective assessment of intermediate quality level of audio systems,” *International Telecommunication Union Radiocommunication Assembly*, 2014.
- [27] Sri Karlapati, Alexis Moinet, Arnaud Joly, Viacheslav Klimkov, Daniel Sáez-Trigueros, and Thomas Drugman, “Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech,” *arXiv preprint arXiv:2004.14617*, 2020.
- [28] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, “World: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
- [29] W Van Summers, David B Pisoni, Robert H Bernacki, Robert I Pedlow, and Michael A Stokes, “Effects of noise on speech production: Acoustic and perceptual analyses,” *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.