



Gradual modeling of the Lombard effect by modifying speaker embeddings from a Text-To-Speech model

Thiago Henrique Gomes Lobato¹, Magnus Schäfer¹

¹HEAD acoustics GmbH

thiago.lobato@head-acoustics.com, magnus.schaefer@head-acoustics.com

Abstract

This work proposes to modify clean speech into Lombard-like speech by modifying speaker embeddings used to condition a text-to-speech model. A feedforward network learns to map embeddings of plain speech to those of Lombard speech using paired data from the Audio-Visual Lombard Grid corpus. Signal level is then increased as per ITU-T P.1150, and a neural vocoder performs time stretching. We show that the resulting speech retains most of the speaker's identity while incorporating relevant Lombard characteristics. Additionally, by properly interpolating embeddings, we propose an approach to gradually model Lombard speech as a function of the background noise level. Listening tests show about a 1.12-point Mean opinion score (MOS) increase in speech plausibility in a loud background context, with only a 0.5-point MOS decrease in speaker similarity compared to an ideal Lombard speech interpolation. Sound samples are available at: <https://github.com/Thiagohgl/interspeech-gradual-lombard>.

Index Terms: speech evaluation, Lombard effect, voice transformation, ITU standard

1. Introduction

The Lombard effect is defined as the involuntary adaptation speakers perform on their speech when strong background noise is present [1][2]. Depending on the speaker, the properties of their voice and speech may drastically change. Thus, having Lombard-compatible samples is essential for a holistic evaluation of communication systems in which speech is in the foreground and strong background noise may be present, such as in-car communication systems or phone-calls in a loud environment. Even though it is possible to use real recordings of Lombard speech for some cases, this could easily become prohibitive with respect to cost and recording time – particularly if different background noise scenarios and a diverse corpus should be considered. Additionally, there are cases where existing signals are already standardized for certain tests [3], and recording additional material is not feasible at all. Thus, a scalable approach to generate arbitrary Lombard speech for a given set of speakers is something of considerable value.

For telecommunication applications, a naïve transformation approach is recommended in ITU-T P.1150 [4], which consists of applying a speech level calibration as a function of the background noise level. While this step does consider one aspect of Lombard speech (i.e. the level), it disregards many other relevant points such as pitch and duration.

The rapid advancement of artificial intelligence (AI) and deep learning techniques has led to significant progress in

various domains of speech processing [5], including speech recognition [6][7], text-to-speech (TTS) synthesis [8][9], speaker identification [10], and more relevant for this work, voice transformation/voice cloning [11][12][13]. Those advances served as motivation to explore deep learning models for the Lombard transformation problem.

The general idea of voice transformation is to take an original speaker and, while retaining the textual information (i.e. the spoken words), change its characteristics either with respect to its emotions or completely changing their identity (i.e. voice cloning). While approaches that apply a direct transformation to the speaker mel-spectrograms have recently shown good results [14], they often demand a high amount of data for pre-training, which is not feasible for the modelling of the Lombard effect, since only limited data is available. In the Hurricane challenge [15] for speech intelligibility, some data with Lombard speech was available and one submission was based on a model-based transformation from plain to Lombard speech (the GlottLombard [16]). It, however, relies on generating a per-speaker dataset of Lombard utterances (not feasibly for zero-shot transformation) and lacks the representation power of modern deep learning speech generators. Approaches such as FreeVC [13] are more promising, since their speaker-embeddings are pre-trained with considerably more data and thus show a high generalization. The idea is to condition the generation to speaker embeddings and text, which effectively considers the voice transformation/cloning problem as a conditioned TTS one. Thus, if we have a decoder capable of generating the desired transformation of our original speech, we just need a good approach to map the embeddings themselves. Indeed, latent space manipulation was already shown to be effective in many applications such as semantic image manipulation [17][18] and even in speech manipulation [19].

Based on this, this work proposes to learn a transformation between normal (i.e. plain) speech embeddings and their Lombard variants and use those to generate Lombard speech. This is done by using a TTS model conditioned on speaker embeddings, and the transformation is learned with a simple feedforward network, since the amount of available Lombard data is very low. One of the main contributions of this work is to show that the difference between plain and Lombard speakers can be considered directly on the speakers' embeddings, and thus, if the TTS is able to represent the speaker, good results can be achieved. Additionally, we propose a method to calibrate latent space interpolations so that we can control the intensity of the Lombard speech for different

background noises. The paper is divided by first introducing the used data and general modeling considerations, then explaining each component in more detail. Afterwards, our methodology for calibrating the interpolation is presented and we explain the listening test used to validate our approach. This is followed by the results and conclusions.

2. Data and Modeling

We show our complete workflow in Figure 1. The main idea is that as long as we can generate Lombard embeddings of input speakers and have a TTS model that can properly leverage those embeddings, Lombard versions of a speaker are possible. These, however, may still miss important properties such as correct level (which is arbitrary in a TTS model) and the artificial longer duration, which is an essential part of Lombard speech [2]. Those can be easily controlled, the first with a simple factor and the second by stretching a spectrogram input for a vocoder, which in this work was the Vocos [20]. The individual parts of the diagram are going to be discussed in the following sections.

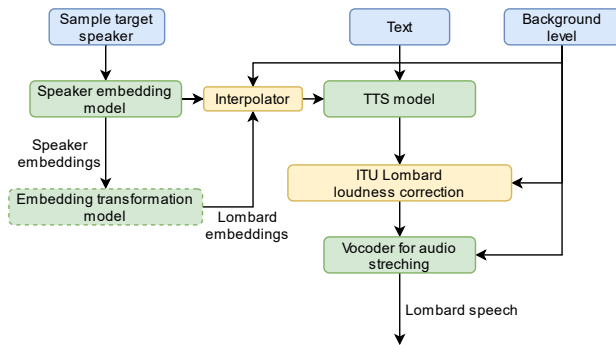


Figure 1: *Workflow of the proposed approach. We transform plain speaker embeddings into Lombard embeddings and then apply level and length corrections.*

2.1. Datasets

2.1.1. Lombard Grid

The Lombard Grid Corpus [21] contains 50 utterance pairs (plain-Lombard) for 54 different speakers. The utterances are very short, with only around 5 seconds duration, and thus the total amount of signal length for each speaker is very limited. To generate the Lombard variation, a background noise following a typical speech spectrum and with 80 dB sound pressure level was played through headphones, the speakers were then asked to recite the utterances to a person they could see.

Even though the total amount of data is relatively low, these utterance pairs are extremely valuable for investigating the relation between plain and Lombard speech, since it allows us to perform a direct comparison. As an example, we show the average change of the speaker’s fundamental frequency, calculated with CREPE [22], in Figure 2. There we see that the Lombard speech has on average a higher pitch than their plain variants, which is consistent with previous research [1].

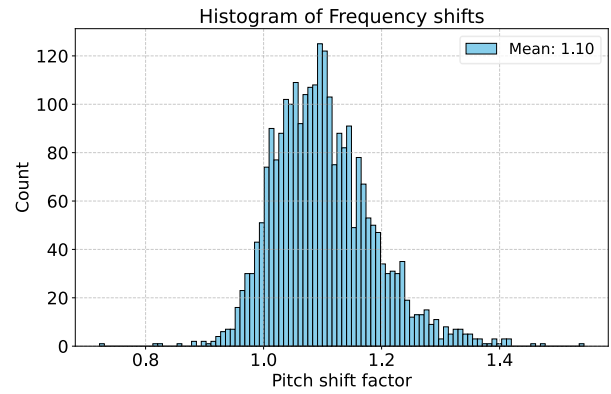


Figure 2: *Pitch shift factor between plain and Lombard speech.*

We also investigate speech duration, obtaining the known result that Lombard Speech is slower than plain speech. The distribution of the Lombard Grid data can be seen in Figure 3.

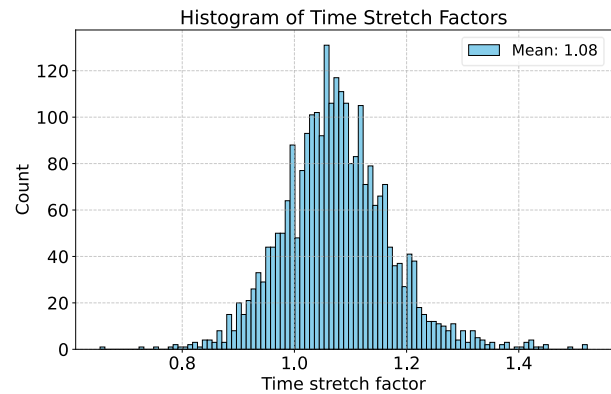


Figure 3: *Time stretch factor between plain and Lombard speech.*

To generate the conditioning signal for the generation of speaker embeddings, we concatenated all sentences for each speaker both for the plain and Lombard case.

2.1.2. ITU-P.501

The ITU-P.501 standard [3] describes different signals that can be used to perform a variety of telecommunication tests. Among them, there’s a sequence of 12 speakers, 6 male and 6 female, for which a Lombard version of them would be especially useful to improve the quality of such tests. Thus, we used them to validate our approach, which is especially interesting since they are extremely different from the signals present on the Lombard Grid dataset.

2.2. TTS model

The TTS model used was the Metavoice 1B¹. This model uses speaker embeddings as conditioning to generate voices with similar identities and was trained for the English language. The speaker embeddings use the same architecture as in [13], being the result of a simple LSTM (Long short-term memory) in a mel-spectrogram representation of the speech signal followed

¹<https://github.com/metavoiceio/metavoice-src>

by a ReLU (Rectified Linear Unit) activation and unity hypersphere projection (unitary norm normalization). It works by predicting Encodec [23] tokens with a GPT (Generative pre-trained transformer) architecture² and then diffuses those tokens in waveforms of different mel-frequency bands with the same approach as in [24]. The generated signal is then filtered with DeepFilterNet [25] to remove artifacts. If a sampling rate higher than 24 kHz is desired, super-resolution approaches such as in [26] could be used.

The TTS model was chosen since its training was focused on emotional speech and, likely because of this, it can generate speech that matches the Lombard variants in the Lombard grid corpus really well (we speculate that Lombard speech characteristics may be somewhat similar to emotive speech like anger, surprise and happiness). This gave us confidence that the model would likely be able to generate believable Lombard speech.

2.3. Embeddings transformation model

The speaker transformation model is a feedforward network composed of two SwiGLU activations, the first with dimensions 256 to 64 and the second 64 to 256, followed by a ReLU activation and projection to the positive unit hypersphere, like the speaker embeddings model.

To train the embedding transformation, we generated plain-Lombard pairs from the Lombard grid corpus and used those as input and output for the model. The model was optimized to maximize the cosine similarity between real and predicted embeddings. For the validation, we used a 5-Fold cross-validation, while the end model used on the ITU P.501 data was trained with the entire dataset.

Since the embeddings are projected on the unit hypersphere, we always project the linear-interpolated embeddings back to it. This guarantees that we do not leave the embeddings manifold.

2.4. Speech stretching

To replicate the additional duration of Lombard speech, we used a speech stretching approach based on a mel-spectrogram vocoder, here Vocos [20]. This approach is similar to the augmentations proposed in [13], but used here to generate the desired signal.

2.5. Interpolating Lombard levels

Lombard speech is not a binary effect. It can not only be present or not present, but it may manifest itself in different levels as a function of the background noise, and thus even though an “on/off” switch is helpful, it disregards many relevant use-cases. There is, however, no database diverse enough to learn a model for different background levels.

To solve this problem, we propose to interpolate the embeddings between plain and Lombard speech. In this way we can generate voices that are “in between” both. The challenge with this approach is to find a suitable relation between the interpolation factor and the background noise. The interpolation function $g(x)$ based on the plain embedding g_{plain} and Lombard embedding g_{lomb} is defined as:

$$g(x) = \text{proj}((1 - 0.01x)g_{plain} + 0.01x \cdot g_{lomb}) \quad (1)$$

In which $\text{proj}(\cdot)$ is the projection operator to the unit sphere and x a “Lombard strength” factor ranging from 0 to 100. The ITU-T P.1150 proposes an increase in level starting at 50 dB background noise, while the background noise at the Lombard grid corpus was 80 dB. Thus, if we consider that 50 dB is plain and 80 dB is Lombard (as given by our transformer), we get the end points.

Based on internal data, we assume that the fundamental frequency changes linearly with the background level. Thus, we estimate the average speaker fundamental frequency for different interpolation factors and find the ideal projection so that it changes linearly with the level. This was done with the ITU speakers to consider the calibration of out-of-distribution samples. The results were projected/smoothed on a cubic fit to reduce the random variations of the estimation. For the speech duration, we use a similar linear relation with a maximum factor of 1.08, which is the average in the Lombard grid corpus. The actual stretching factor is calculated by considering the average duration of the ITU-T P.501 generated signals for each background level. We use this approach to generate equivalent ITU P.501 Lombard-data for 50 dB, 60 dB, 70 dB and 80 dB

3. Listening test

We can easily verify if the generated signals show a similar distribution of fundamental frequency change, level and duration (the last 2 by design). However, it is neither clear if the speakers are consistent throughout the transformations nor if they sound natural in an environment with strong background noise. To check those, we defined a listening test in which we played generations of the following variants: ITU-T P.501 speakers with Lombard transform and only with the ITU-T P.1150 level change recommendation, Lombard Grid speakers with perfect Lombard embeddings and completely random Lombard Grid speakers with the Lombard transform (serving as a lower anchor on speaker similarity). For each speaker, a sequence of generations with increasing background noise (from 50 dB to 80 dB) was played. The background noise was, for each single sentence, a random sample from a restaurant recording, which we selected since it represents a situation in which people may be used to hearing Lombard speech. For each sequence, we evaluated on a scale from 1 to 5:

- Speaker similarity
- Speaker naturalness
- Plausibility of speech progression with increasing background

In total we had 48 comparisons, each around 15 s (please see paper website for examples). The duration of the test was, on average, 30 minutes. The number of participants was 10 with ages ranging from 22 to 46 years.

4. Results

4.1. Embedding transform

The embedding transform obtained a 5-Fold average cosine similarity in the validation set of 0.95. This value is

² Radford, A, Wu, J, Child, R, Luan, D, Amodei, D, Sutskever, I. "Language Models are Unsupervised Multitask Learners", 2019

considerably high, indicating that the plain-Lombard relation can be identified well in the speaker embedding space, at least for the Lombard Grid corpus. The actual evaluation of the embedding model, however, was done in the listening test at the unseen ITU dataset.

4.2. Ideal Interpolation factors

The relation between average fundamental frequency and interpolation strength (defined in $[0,100]$) can be seen in Figure 4, in which each green dot represents a 5 dB step starting from 50 dB background noise:

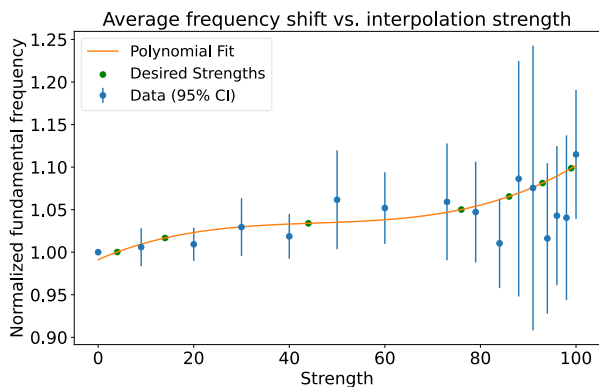


Figure 4: Relation between fundamental frequency and interpolation strength.

The final fundamental shift factor of the Lombard transform is somewhat larger than the value from the Lombard Grid corpus (1.14 instead 1.10). However, since we only have 12 speakers in the ITU dataset and their sentences are considerably more elaborate, we don’t regard this difference as problematic (especially considering the 95% confidence interval).

4.3. Listening test results

The listening test results can be seen in Table 1 for the Lombard Grid ideal interpolation (LG-II), ITU full transform (ITU-FT), ITU level only (ITU-LO) and random speakers (RS). There, we verify that the speaker similarity of ITU transformation is considerably better than random speakers (at about 1.3 points) while only slightly worse (about 0.5 points) than if using real Lombard speech embeddings. This indicates that our approach is only slightly worse than an “ideal” Lombard interpolation with respect to the speaker identity. On the other hand, it is considerably more plausible (1.12 points) than the current ITU level-only approach, indicating that people judge a Lombard speech as more plausible when additional aspects such as pitch and duration are considered, even if the speaker identity may be slightly worse than the ideal interpolation.

We also verify that the approaches with transformations have a slightly better naturalness than the ITU level-only variation, indicating that the original signals may not be ideal in the loud background. For the overall quality, all approaches have very similar values, with a very modest maximum at the ITU variant. This implies that our transformation has no significant effect on the sound quality of the original text-to-speak model.

Table 1: MOS results of the listening test. Blue is our transformation approach in the out-of-training-distribution ITU data.

| Case | Speaker similarity | Naturalness | Plausibility | Overall Quality |
|----------|--------------------|-----------------|-----------------|-----------------|
| LG – II | 3.81 \pm 0.21 | 3.19 \pm 0.2 | 3.08 \pm 0.2 | 3.06 \pm 0.18 |
| ITU – FT | 3.26 \pm 0.22 | 3.47 \pm 0.17 | 3.31 \pm 0.19 | 3.12 \pm 0.17 |
| ITU- LO | 4.86 \pm 0.07 | 3.08 \pm 0.16 | 2.19 \pm 0.18 | 3.37 \pm 0.16 |
| RS | 1.88 \pm 0.17 | 2.92 \pm 0.18 | 2.48 \pm 0.16 | 2.88 \pm 0.17 |

By performing some interviews after the test, the biggest complaint of participants with respect to the quality was that the last example in the sequence with the loudest background noise had a too quiet speech. This may indicate that the current level change on the ITU-T P.1150 standard may be too small for real backgrounds. However, in very loud environments, people usually get closer to each other to speak, so that we may associate it to a higher level than the ones people are actually producing due to the smaller distance to the speaker. This is something that could be investigated in future work.

5. Conclusions

We’ve presented a method to generate Lombard speech for different speakers in a scalable fashion. This was done by transforming plain speaker embeddings into their Lombard variants and using those as conditioning to a TTS model. The generated speech signals are then post-processed with standardized level amplifications and voice stretching using a Vocoder in order to match the average data distribution of the Lombard Grid dataset. Our Lombard transformations show a considerable improvement in their plausibility at loud backgrounds (about 1.12 points in MOS score) while being judged only slightly worse to ideal Lombard speech signals with respect to the speaker similarity (about 0.5 points in MOS score).

Additionally, we showed that it is possible to map the strength of a latent space interpolation to different background noise levels, which allows the generation of Lombard speech signals at different levels. This is a very relevant aspect as different background noise levels need to be considered for many application scenarios. Thus, we believe the approach proposed in this paper could be readily applied to, e.g., generate more realistic Lombard signals for telecommunication testing, which is reinforced by the fact that we used sounds from the ITU-T P.501 standard in our tests.

Future research could deal with the generation of Lombard speech in a dynamic varying background, for example when telephoning in a busy street where cars appear randomly at the street. An investigation of the actual generated/expected loudness change of the Lombard speech could also be done, since it was the main complaint of the participants during the test. Another interesting point could be to apply the Lombard conversion to other types of TTS models that may have a higher expressivity or cover different languages besides English.

6. References

- [1] J. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.* vol. 93, no. 1, pp. 510–524, 1993.
- [2] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3261–3275, 2008.
- [3] International Telecommunication Union. Test signals for use in telephony (ITU-T Recommendation P.501). Geneva: ITU; 2012.
- [4] International Telecommunication Union. ITU-T P.1150: Requirements for conversational interactive voice response systems. Geneva: ITU; 2019.
- [5] A. Mehrish, N. Majumder, R. Bhardwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, 2023.
- [6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023.
- [7] C. Yu, M. Kang, Y. Chen, J. Wu, and X. Zhao, "Acoustic modeling based on deep learning for low-resource speech recognition: An overview," *IEEE Access*, vol. 8, pp. 163829–163843, 2020.
- [8] S. E. Eskimez et al., "E2 TTS: Embarrassingly Easy Fully Non-Autoregressive Zero-Shot TTS," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, Macao, 2024, pp. 682–689, doi: 10.1109/SLT61566.2024.10832320.
- [9] H. Barakat, O. Turk, and C. Demiroglu, "Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources," *J. Audio Speech Music Process.*, 2024, doi: 10.1186/s13636-024-00329-7
- [10] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities," *IEEE Access*, vol. 9, pp. 79236–79263, 2021, doi: 10.1109/ACCESS.2021.3084299.
- [11] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance," in *Proc. Interspeech 2018*.
- [12] A. Baade, P. Puyuan, and D. Harwath, "Neural codec language models for disentangled and textless voice conversion," in *Proc. Interspeech 2024*. ISCA, 2024.
- [13] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10095191.
- [14] T. Kaneko, H. Kameoka, K. Tanaka, and Y. Kondo, "FastVoiceGrad: One-step diffusion-based voice conversion with adversarial conditional diffusion distillation," in *Proc. Interspeech 2024*, 2024, pp. 192–196, doi: 10.21437/Interspeech.2024-2387.
- [15] Cooke, M., Mayo, C., Valentini-Botinhao, C. (2013) Intelligibility-enhancing speech modifications: the hurricane challenge. *Proc. Interspeech 2013*, 3552-3556, doi: 10.21437/Interspeech.2013-764.
- [16] Suni, A., Karhila, R., Raitio, T., Kurimo, M., Vainio, M., Alku, P. (2013) Lombard modified text-to-speech synthesis for improved intelligibility: submission for the hurricane challenge 2013. *Proc. Interspeech 2013*, 3562-3566, doi: 10.21437/Interspeech.2013-766.
- [17] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Advances in Neural Information Processing Systems*, vol. 31, pp. 10215–10224, 2018.
- [18] P. Zhuang, O. Koyejo, and A. G. Schwing, "Enjoy Your Editing: Controllable GANs for Image Editing via Latent Space Navigation". in *Proc. 9th Int. Conf. Learn. Representations (ICLR)*, 2021.
- [19] S. Jo, Y. Lee, Y. Shin, Y. Hwang, and T. Kim, "Cross-speaker emotion transfer by manipulating speech style latents," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, doi: 10.1109/ICASSP49357.2023.10095619.
- [20] S. Hubert, "Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024.
- [21] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. Brown, "A corpus of audio-visual Lombard speech with frontal and profile views," *J. Acoust. Soc. Am.*, vol. 143, no. 6, p. EL523, 2018.
- [22] J. Kim, J. Salamon, P. Li, and J. Bello, "CREPE: A convolutional representation for pitch estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2018.
- [23] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Trans. Mach. Learn. Res.*, 2023.
- [24] R. San Roman, Y. Adi, A. Deleforge, R. Serizel, G. Synnaeve, and A. Défossez, "From discrete tokens to high-fidelity audio using multi-band diffusion," in *Proc. 37th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2023.
- [25] H. Schröter, N. Alberto, B. Escalante, T. Rosenkranz, and A. Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [26] H. Liu, C. Ke, T. Qiao, W. Tian, M. D. Wang, and M. Plumbley, "AudioSR: Versatile audio super-resolution at scale," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Republic of Korea, 2024, pp. 1076–1080, doi: 10.1109/ICASSP48485.2024.10447246.