

IMPACT OF NORMALIZATION IN MULTIMODAL AI

Neeraj Kumar



**BHARTI SCHOOL OF TELECOMMUNICATION TECHNOLOGY &
MANAGEMENT**

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2024

©Indian Institute of Technology Delhi - 2024
All rights reserved.

IMPACT OF NORMALIZATION IN MULTIMODAL AI

by

Neeraj Kumar

Bharti School of Telecommunication Technology & Management

Submitted

in fulfillment of the requirements of the degree of **Doctor of Philosophy**

to the



Indian Institute of Technology Delhi

JULY 2024

Certificate

This is to certify that the thesis titled “**IMPACT OF NORMALIZATION IN MULTIMODAL AI**” being submitted by **NEERAJ KUMAR** for the award of **Doctor of Philosophy** in **Bharti School of Telecommunication Technology & Management**, is a record of bona fide work carried out by him under my guidance and supervision at the **Bharti School of Telecommunication Technology & Management, Indian Institute of Technology Delhi**. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Ankur Narang

Ankur Narang
VP of AI
Hike Private Limited
New Delhi- 110016

Brejesh Lall
Professor
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi- 110016

Acknowledgements

I owe my ability to pursue my PhD to the unwavering support of my family, especially my parents, who have shown incredible understanding and made significant sacrifices throughout their lives. I am also immensely grateful to my wife, Shinjini, for her constant presence and support. It was the nurturing and supportive environment at home that allowed me to dedicate the majority of my waking hours to my academic pursuits.

I extend my deepest appreciation to my advisors, Prof. Brejesh Lall and Dr. Ankur Narang. Your faith in me and the freedom you granted me to explore topics of personal interest have been invaluable. I am thankful for the continuous encouragement, feedback, guidance, and support you provided, serving as a constant source of inspiration.

Throughout my doctoral journey, I held a full-time position at Hike Private Limited, a workplace I consider exceptional. I have gained a wealth of knowledge from my colleagues, and my PhD would not have been achievable without the unwavering support of my friends and coworkers at Hike. I express my gratitude to my colleagues and teammates, including Srishti, Pranshu, Dipankar, and Kush, for their invaluable support and technical insights.

I would like to acknowledge all my friends at IITD for their technical discussions, which greatly contributed to my research and enhanced my comprehension. Additionally, they made life enjoyable and thought-provoking with their shared experiences.

A special mention goes to my remarkable circle of friends, including Sohil, Pallav, Mohit, Ayan, Kaushal, Jayant and Vikas for the joy, humor, and laughter they brought into my life.

Lastly, I am thankful to God for everything and all that I have achieved.

Neeraj Kumar

Abstract

Multimodal deep learning systems, which leverage multiple modalities such as text, image, audio, and video, have demonstrated superior performance compared to single-modality systems. The process of multimodal machine learning encompasses several stages, including representation, translation, alignment, fusion, and co-learning.

Three distinct fusion techniques are examined: early fusion, intermediate fusion, and late fusion. Early fusion involves amalgamating raw data from diverse modalities into a unified representation before the learning phase. Intermediate fusion entails transforming raw inputs into a higher-level representation through a series of layers. Late fusion, on the other hand, entails merging predictions from each modality to arrive at a final decision. Various fusion methods, including element-wise summation, weighted averaging, bilinear product, rank minimization, and attention mechanisms, have been proposed previously.

However, existing methods face several challenges. These include discrepancies or misaligned spatial dimensions of intermediate features from different modalities, overfitting issues in late fusion networks due to the increased parameter count, and complications in utilizing pre-trained weights from uni-modal networks when incorporating intermediate fusion. There is also the potential for performance degradation in multimodal models compared to uni-modal ones when employing a joint training strategy, which contradicts the goal of enhancing performance through multimodal integration. Different modalities may converge at varying rates, resulting in uncoordinated convergence issues.

Our research has delved into the role of normalization in various multimodal applications, such as audio-to-video synthesis, text-to-speech synthesis, and uni-modal feature extraction. We have introduced innovative normalization techniques that have proven effective across different multimodal applications. Our approach, known as multimodal normalization, is designed to capture the interdependence between different domains by utilizing a mixture of multivariate Gaussian distributions. This method helps prevent the model from experiencing mode collapse, a common issue when trying to model data distribution solely with multivariate Gaussian

distributions. The integration of multimodal normalization has yielded outstanding results in our proposed video synthesis, involving the transformation of audio and a single image, outperforming previous methods on various metrics such as SSIM (structural similarity index), PSNR (peak signal-to-noise ratio), CPBD (image sharpness), WER (word error rate), blinks per second, and LMD (landmark distance). In the context of video emotion detection, we harnessed the power of the mel-spectrogram and optical flow within the multimodal normalization approach to learn affine parameters, thereby contributing to enhanced accuracy.

We have introduced a novel normalization framework for text-to-speech applications, which is designed to capture stylistic features at both the speaker and frame levels, in addition to accommodating style-agnostic features. This is achieved through the utilization of learnable parameters denoted as γ (scale) and β (bias) within the normalization framework. To compute the various learnable parameters associated with speaker embeddings from the speaker encoder, pitch, and energy, we have proposed two distinct approaches: one based on convolutional networks and the other employing a multi-head attention network. We establish the effectiveness of our proposed architecture through comprehensive evaluations on the VCTK and LibriTTS datasets. Our assessments include visualizing the Hessian matrix of the proposed model, employing various quantitative metrics to measure speech distortion, evaluating Mean Opinion Scores (MOS), and conducting an in-depth analysis of speaker embeddings generated by our novel speaker encoder model.

We introduce a novel approach known as Kullback-Leibler (KL) Regularized Normalization (KL-Norm). This technique enhances the stability of normalized data, thereby contributing to improved generalization by reducing overfitting. It generalises well in adapting to out-of-domain distributions while eliminating irrelevant biases and features, all accomplished with minimal impact on model parameters and memory resources. Our extensive empirical assessments across various low-resource natural language processing (NLP) and speech-related tasks substantiate the superior performance of KL-Norm when compared to well-known normalization and regularization methods.

Lastly, Our contributions encompass the introduction of innovative multimodal normalization techniques, which have proven advantageous in diverse applications such as speech-driven video generation and video emotion detection, the development of a normalization framework within text-to-speech systems to capture both stylistic and style-agnostic features, and the introduction of KL Regularized normalization for addressing low-resource natural language processing challenges. This research topic has resulted in the creation of novel normalization architectures and an extensive exploration of the impact of normalization techniques across a wide range of multimodal applications.

सार

मल्टीमॉडल डीप लर्निंग सिस्टम, जो पाठ, छवि, ध्वनि और वीडियो का उपयोग करते हैं, एकल-मॉडलिटी सिस्टम की तुलना में बेहतर प्रदर्शन करते हैं। मल्टीमॉडल मशीन लर्निंग में प्रतिनिधित्व, अनुवाद, संरेखण, संयोजन और सह-अधिगम जैसे चरण शामिल हैं। संयोजन तकनीकों में प्रारंभिक, मध्यवर्ती और अंतिम संयोजन शामिल हैं। प्रारंभिक संयोजन में विभिन्न माध्यमों के कच्चे डेटा को एकीकृत प्रतिनिधित्व में मिलाया जाता है। मध्यवर्ती संयोजन में कच्चे इनपुट को परतों के माध्यम से उच्च-स्तरीय प्रतिनिधित्व में बदल दिया जाता है। अंतिम संयोजन में प्रत्येक माध्यम से प्राप्त पूर्वानुमानों को अंतिम निर्णय में मिलाया जाता है। संयोजन विधियों में तत्व-वार जोड़, भारित औसत, द्विखंडित उत्पाद, रैंक न्यूनतमकरण और ध्यान तंत्र शामिल हैं। चुनौतियों में विभिन्न माध्यमों के मध्यवर्ती विशेषताओं के विसंगत या असंगत स्थानिक आयाम, अंतिम संयोजन नेटवर्क में अधिक अभ्यस्तता के मुद्दे और मध्यवर्ती संयोजन में पूर्व-प्रशिक्षित भार का उपयोग करने में कठिनाइयाँ शामिल हैं। संयुक्त प्रशिक्षण के दौरान विभिन्न माध्यमों के असंयोजित समापन दर के कारण मल्टीमॉडल मॉडल का प्रदर्शन एकल-मॉडल की तुलना में खराब हो सकता है।

हमारा शोध ध्वनि-से-वीडियो निर्माण, पाठ-से-भाषण निर्माण और एकल-माध्यम विशेषता निष्कर्षण जैसे मल्टीमॉडल अनुप्रयोगों में सामान्यीकरण का पता लगाता है। हम विभिन्न क्षेत्रों के बीच परस्पर निर्भरता को पकड़ने के लिए बहुसांख्यिक गॉसियन वितरणों के मिश्रण का उपयोग करके मल्टीमॉडल सामान्यीकरण प्रस्तुत करते हैं, जो मोड कोलैप्स को रोकता है। यह वीडियो निर्माण प्रदर्शन में सुधार करता है, जैसा कि सरचनात्मक समानता सूचकांक, पीक सिग्नल-टू-शोर अनुपात, छवि तीक्ष्णता, शब्द त्रुटि दर, प्रति सेकंड झपकी और लैंडमार्क दूरी जैसे मापदंडों द्वारा प्रदर्शित किया गया है। पाठ-से-भाषण के लिए, हम एक सामान्यीकरण ढांचा प्रस्तुत करते हैं जो वक्ता और फ्रेम-स्तरीय शैलीगत विशेषताओं को पकड़ता है, जो गामा और बीटा जैसे सीखने योग्य मापदंडों का उपयोग करता है। हम इन मापदंडों की गणना के लिए संपूर्ण नेटवर्क और बहु-प्रमुख ध्यान नेटवर्क का प्रस्ताव रखते हैं। वीसीटीके और लिब्रिट्टीएस डेटासेट पर हमारे आर्किटेक्चर की प्रभावशीलता हेसियन मैट्रिक्स, भाषण विकृति मापदंड, औसत राय स्कोर और वक्ता एम्बेडिंग द्वारा मापी गई है। हम कुलबैक-लेबलर (केएल) नियमितीकृत सामान्यीकरण भी प्रस्तुत करते हैं, जो डेटा स्थिरता को बढ़ाता है और अधिक अभ्यस्तता को कम करते हुए सामान्यीकरण में सुधार करता है। केएल सामान्यीकरण बाहरी-डोमेन वितरणों के अनुकूलन में अच्छा प्रदर्शन करता है और अप्रासंगिक पूर्वाग्रह और विशेषताओं को समाप्त करता है, सभी न्यूनतम मापदंड और स्मृति प्रभाव के साथ। विभिन्न निम्न-संसाधन प्राकृतिक भाषा संसाधन और भाषण कार्यों में केएल सामान्यीकरण का उत्कृष्ट प्रदर्शन हमारे व्यापक अनुभवजन्य मूल्यांकनों द्वारा सिद्ध है।

हमारे योगदान में भाषण-चालित वीडियो निर्माण और वीडियो भाव पहचान के लिए नवाचारी मल्टीमॉडल सामान्यीकरण तकनीकें, पाठ-से-भाषण के भीतर शैलीगत और शैली-अज्ञेय विशेषताओं को पकड़ने वाला सामान्यीकरण ढांचा, और निम्न-संसाधन प्राकृतिक भाषा संसाधन चुनौतियों के लिए केएल नियमितीकृत सामान्यीकरण शामिल हैं। इस शोध ने सामान्यीकरण आर्किटेक्चर में नवाचार किया है और विभिन्न मल्टीमॉडल अनुप्रयोगों में सामान्यीकरण तकनीकों के प्रभाव का व्यापक रूप से अन्वेषण किया है।

Contents

Certificate	i
Acknowledgements	iii
Abstract	v
List of Figures	xix
List of Tables	xxiv
1 Introduction	1
1.1 Motivation	1
1.2 Contribution	3
1.2.1 Audio to Video Synthesis	4
1.2.2 Text to Speech Synthesis	5
1.2.3 Normalization in Low Resource Tasks	6
1.3 Thesis Outline	7
2 Background & Related Work	9
2.1 Normalization	10

2.2	Multimodal fusion methods	14
2.3	Audio to Video Generation	16
2.3.1	Robust One shot Audio to Video Generation	16
2.3.2	One Shot Audio to Animated Video Generation	19
2.3.3	MultiModal Normalization	20
2.4	Text To Speech	21
2.4.1	Zero-shot Normalization Driven Multi-Speaker Text to Speech Synthesis	22
2.4.2	Style Description based Text-to-Speech with Proportional Prosodic Layer Normalization based Diffusion GAN	25
2.5	KL Regularized Normalization Framework for Low Resource NLP Tasks	26
2.5.1	Low resource NLP and Speech	26
2.5.2	KL Regularization	27
3	Theoretical Foundation	29
3.1	Multi Modal Normalization	30
3.1.1	Introduction	30
3.1.2	Multi-Modal Normalization - Theory	30
3.2	Normalization in Text to Speech	32
3.2.1	Introduction	32
3.2.2	Convolution based Normalization	32
3.2.3	Multi Head Attention Based Normalization	33
3.2.4	Normalization Analysis with Hessian eigenvalue density function	35
3.3	KL Regularized Normalization	37
3.3.1	Introduction	37

3.3.2	Preliminaries : Batch Normalization	37
3.3.3	KL Regularized Batch Normalization	39
4	Audio to Video Synthesis	45
4.1	Robust One shot Audio to Video Synthesis	46
4.1.1	Introduction	46
4.1.2	Architectural Design	47
4.1.3	Losses	51
4.1.4	Datasets and Training	54
4.1.5	Evaluation Metrics	55
4.1.6	Experiments and Results	56
4.2	One Shot Audio to Animated Video Generation	61
4.2.1	Introduction	61
4.2.2	Architectural Design	62
4.2.3	Losses	64
4.2.4	Datasets & Training	66
4.2.5	Evaluation Metrics	67
4.2.6	Experiments and Results	68
4.3	MultiModal Normalization	70
4.3.1	Introduction	70
4.3.2	Architectural Design - Multi Modal Normalization in Audio to Video Generation	72
4.3.3	Architecture Design- Multi Modal Normalization in Video Emotion Detection	78

4.3.4	Losses	79
4.3.5	Datasets and Training	81
4.3.6	Evaluation Metrics	83
4.3.7	Experiments and Results	85
4.4	Summary	93
5	Text To Speech Synthesis	95
5.1	Zero-shot Normalization Driven Multi-Speaker Text to Speech Synthesis	96
5.1.1	Introduction	96
5.1.2	Architectural Design	98
5.1.3	Losses	104
5.1.4	Datasets and Training	105
5.1.5	Evaluation Metrics	107
5.1.6	Experiments and Results	108
5.1.7	Extension of Proposed Method	118
5.2	Style Description based Text-to-Speech with Proportional Prosodic Layer Normalization based Diffusion GAN	122
5.2.1	Introduction	122
5.2.2	Architectural Design	123
5.2.3	Losses	127
5.2.4	Datasets and Training	128
5.2.5	Experiments and Results	130
5.2.6	Speech Quality	130
5.2.7	Qualitative Results	131

5.3	Summary	132
6	Normalization Framework for Low Resource Tasks	135
6.1	KL Regularized Normalization	135
6.1.1	Introduction	135
6.1.2	Algorithm	138
6.1.3	Training	142
6.1.4	Benchmarking	143
6.1.5	Ablation Study	148
6.2	Summary	152
7	Conclusion and Future Work	153
7.1	Conclusion	153
7.2	Scope of Future Work	154
	Bibliography	159
	List of Publications	189
	Biography	191

List of Figures

2.1	Illustration of normalization operations discussed in this paper.	11
3.1	Higher level architecture of Multi Modal Normalization	30
3.2	Convolution based normalization in proposed ZSM-SS architecture	33
3.3	Multi-Head Attention Based Normalization in proposed ZSM-SS architecture .	34
3.4	The KL Normalized distribution(black contour) tries to match the fixed prior and seeing a hole.	41
3.5	The KL Normalized distribution(pink contour) tries to match the learnable prior(black contour) and is modified to fit the prior.	42
4.1	Model for generating robust and high-quality videos. This uses deep speech audio features to be fed into SPADE Generator and 2 discriminators i.e frame discriminator which is a multi-scale discriminator for frame generation and another discriminator for better lip synchronization.	48
4.2	the SyncNet architecture for better lip synchronization which is trained on GRID dataset with contrastive loss and then used its loss in our proposed architecture	48
4.3	6 landmark points for blink loss	53
4.4	(a) Top: Female uttering the word “now”(b) Bottom: Male uttering the word “bin”	56
4.5	(a) Top: Movement of eyes while speaking (b) Bottom: Speaker uttering a hindi male name “Modi”	57

4.6	(a) Top: Speaker uttering the word “Please”on the GRID dataset (b) Bottom: Speaker uttering the word “Please”on the LOMBARD GRID dataset	57
4.7	Distribution of user scores for the online Turing test	59
4.8	(a) Top: Stage 1 of OneShotAu2AV with a generator and three discriminators for generating human-domain video. (b) Bottom: Stage 2 of OneShotAu2AV with a generator, temporal predictor and a discriminator for generating a high quality animated video.	63
4.9	Unet architecture	64
4.10	Left side: Animated output speaking “now”; Right side: Head movement of female anime.	68
4.11	Left side: Animated output with eye blinks; Right side: Eyebrow movements of male while speaking.	68
4.12	Left side: Anime speaking the Hindi word “modi”; Right side: Anime speaking the Bengali word “aache”.	69
4.13	Proposed architecture for Audio to Video synthesis- MultiNormA2V	72
4.14	Multimodal normalization residual architecture	73
4.15	Class activation map layer architecture in generator	73
4.16	Generator architecture	73
4.17	Architecture to calculate the affine parameters from video features in multimodal normalization	74
4.18	Architecture to calculate the affine parameters from audio features in multimodal normalization	74
4.19	Keypoint heatmap predicted architecture	75
4.20	Optical flow predictor architecture	76
4.21	Discriminator architecture	78
4.22	Emotion Detection Architecture	78

4.23	Top: The speaker speaking the word “bin”, Middle : The speaker speaking the word “please”, Bottom: The speaker blinking his eyes	87
4.24	Top: Actual frames of voxceleb2[1] data set , Middle : Predicted frames from proposed method, Bottom: Predicted frame from [2]	88
4.25	Top: The speaker with different expressions, Middle1 : CAM based attention map, Middle2: Predicted optical flow from the optical flow generator architecture, Bottom: Predicted Key-points from Key-point predictor architecture	89
4.26	A blink event is identified at the point where a noticeable decrease in the EAR signal is observed (marked by the blue dot). We define the initiation of the blink as indicated by the green dot and the termination of the blink by the red dot, both aligned with the peaks on either side of the blink location (refer to the color figure online).	90
4.27	Top: Actual frames of speaker of GRID data set. Middle: Predicted frames from proposed method with keypoints predicted from keypoint predictor. Bottom: Predicted frames from the FOMM method [3]	91
4.28	Distribution of user scores for the online Turing test	92
5.1	ZSM-SS : Zero-Shot style based text to Speech Generation Architecture	99
5.2	Left: FFT block , Centre : Variance Adaptor[4] , Right : Variance Predictor[4] .	100
5.3	Speaker Encoder	100
5.4	Convolution based normalization in proposed ZSM-SS architecture	102
5.5	Multi-Head Attention Based Normalization in proposed ZSM-SS architecture .	103
5.6	Spectral densities of ZSM-SS Model without normalization at time step, $t = 120000$ and have used $\sigma^2 = 10^{-5}$, $m = 90$ for our experiments.	109
5.7	Spectral densities of ZSM-SS Model with normalization at time step, $t = 120000$	109
5.8	Skewness of ZSM-SS Model.	110
5.9	Kurtosis of ZSM-SS Model.	110

5.10 t-SNE visualization of speaker embeddings generated from speaker encoder model. Cluster id 0 to 5 refers to female speakers and 6 to 10 refers to male speakers	111
5.11 Left: Cross Similarity, Right: Normalized histogram of similarity values between utterances of actual speaker (x-axis) and generated speaker (y-axis) of VCTK dataset	111
5.12 t-SNE visualization of speaker embeddings of generated samples of VCTK dataset. Cluster id 0 to 5 refers to female speakers and 6 to 10 refers to male speakers	113
5.13 t-SNE Visualization of speaker embeddings of male actual and generated samples of VCTK dataset. The left side shows the actual embedding space of all male speakers in the dataset. Right side shows the embedding space of train(green), val(red: cluster id - 7,26,42) and test(black: cluster id - 0,6,9,12,13)	113
5.14 Cross Similarity between utterances of actual speaker (x-axis) and generated speaker (y-axis) of VCTK dataset	114
5.15 Normalized histogram of similarity values between utterances actual speaker and generated speaker on VCTK dataset	115
5.16 Left: Synthesized samples on a reference speaker and text, Centre: Pitch is modulated by increasing the F0 to 1.25F0 keeping the energy values constant on same reference speaker and text Right: Energy values are reduced from E to 0.5E keeping the pitch values constant on same reference speaker and text . . .	119
5.17 Probability of belonging to a speaker class on few shot approach with different number of unseen speaker samples and text pairs.	121
5.18 Prosodic Diff-TTS architecture	124
5.19 Left: Multi Head Attention block , Centre : Variance Adapter[4] , Right : Variance Predictor	125
5.20 Style token Generator using pretrained BERT	126
5.21 Proportional Prosodic Layer Normalization architecture	127
5.22 Discriminator architecture of Prosodic Diff-TTS	127

5.23	Plot of energy and pitch of synthesized samples with real speech(Content: Your memory must be conveniently short, chafed the master , Style: Please say a loud girl with a bass)	131
5.24	Plot of mel-spectrogram of synthesized samples with real speech. (Content: Your memory must be conveniently short, chafed the master , Style: Please say a loud girl with a bass)	132
6.1	Architecture of proposed KL Regularized normalization framework	138
6.2	Losses of KL-Norm were observed across different values of β	147
6.3	Accuracy on fixed and learnable prior on MNLI dataset.	149
6.4	Accuracy on fixed and learnable prior on SNLI dataset.	149
6.5	Accuracy on fixed and learnable prior on Google Command dataset.	150
6.6	Effect of control rate and Batch Size on MNLI dataset	150
6.7	Effect of epochs on MNLI dataset	151

List of Tables

4.1	Comparison of OneShotA2V with RSDGAN and Speech2Vid for SSIM, PSNR and CPBD	58
4.2	The above comparison is on lip synchronizing metric i.e word error rate (WER) and average content distance (ACD) by calculating cosine distance (ACD-C) and euclidean distance (ACD-E) between the actual image and the generated image.	58
4.3	Psychophysical Evaluation (in percentages) based on users rating	58
4.4	Ablation Study on the GRID dataset where, CL is the contrastive loss ,TAL is the multi-scale temporal adversarial loss and BL is the Blink loss	60
4.5	Ablation Study on the LOMBARD GRID dataset where, CL is the contrastive loss, TAL is the multi-scale temporal adversarial loss and BL is the Blink loss	60
4.6	Comparison of OneShotAu2AV with U-GAT-IT and RecycleGAN	69
4.7	Comparison of OneShotAu2AV with U-GAT-IT and RecycleGAN for KID, WER and Blink/sec.	69
4.8	Ablation Study of Stage 2 on the GRID dataset where, Base model is U-GAT-IT architecture, RL is the recycle loss, LSL is the lip synchronisation loss loss and BL is the Blink loss	70
4.9	Psychophysical Evaluation (in percentages) based on users rating	70
4.10	Comparison of the proposed method(MultiNormA2V-keypoint and MultiNormA2V-optical) with other previous works for GRID data set	85

4.11 Comparison of the proposed method(MultiNormA2V) with other previous works for CREMA-D data set	86
4.12 Comparison of the proposed method(MultiNormA2V-keypoint and MultiNormA2V-optical) with other previous works for GRID Lombard data set	86
4.13 Comparison of the proposed method(MultiNormA2V-keypoint and MultiNormA2V-optical) with other previous works for VOXCELEB2 data set	86
4.14 Ablation study of different networks of multimodal normalization on GRID data set	90
4.15 Incremental study of Multi Modal Normalization on Grid dataset	91
4.16 Psychophysical evaluation (in percentages) based on users rating on GRID dataset	92
4.17 Adding multi modal normalization increases the accuracy	93
5.1 Comparison of XLSR based proposed speaker encoder against pretrained Wav2vec2.0 and Deep Speech 2	112
5.2 MOS and SMOS score on ZSM-SS with 95% confidence interval for VCTK and LibriTTS dataset. GT - ground Truth , GTmel - ground truth mel-spectrogram with waveglow as vocoder, Conv - Convolution based normalization in ZSM-SS with waveglow vocoder , Attention - Multi head attention based normalization with waveglow vocoder in ZSM-SS with waveglow architecture	116
5.3 Quantitative metrics on ZSM-SS on zero-shot approach. Conv: Convolution based normalization in ZSM-SS , Attention : Multi head attention based normalization in ZSM-SS, 1- VCTK dataset, 2- LibriTTS dataset	117
5.4 Ablation Study of ZSM-SS. BM is Base Model without normalization method. SE is speaker embedding in normalization, P and E are pitch and energy values in the normalization network.	117
5.5 Mean Opinion Score for the naturalness(N) and similarity(S) of ZSM-SS. BM is Base Model without normalization method. SE is speaker embedding in normalization, P and E are pitch and energy values in the normalization network.	118

5.6	Metric for zero-shot approach for convolution based normalization in ZSM-SS for VCTK dataset. Ratio-range denotes the range of ratio of reference speech length with target speech length	118
5.7	Root Mean square error loss by doing the pitch , Energy and speaker embedding modulation	119
5.8	Metric for few shot approach for convolution based normalization in ZSM-SS for VCTK dataset	120
5.9	Metric for few shot approach for attention based normalization in ZSM-SS for VCTK dataset	120
5.10	MOS and SMOS score on ZSM-SS with 95% confidence interval for VCTK in few shot approach(5 samples). GT - ground Truth , GTmel - ground truth mel-spectrogram with waveglow as vocoder, Conv - Convolution based normalization in ZSM-SS with waveglow vocoder	121
5.11	The accuracy (%) of PromptTTS and Prosodic Diff-TTS on 1-PromptSpeech and 2-LibriTTS datasets.	130
5.12	MOS score of speech quality with 95% confidence intervals.	131
5.13	MOS score of of speech quality at different timesteps with 95% confidence intervals	132
6.1	Average Accuracy and standard deviation of KL-Norm with BERT-base against prior works. Δ are absolute differences with BERT _{Base}	140
6.2	Accuracies of KL-Norm with BERT-base across different training data sizes (200, 400, 600, 800, and 1000 samples). Δ are absolute differences with BERT _{Base} . 141	141
6.3	Accuracies of KL-Norm model with Wav2Vec-base across different training data sizes (300, 600, 900, 1200 and 1500 samples). Δ are absolute differences with Wav2Vec _{Base}	144
6.4	Evaluation of model performance when transferring to novel target datasets. These models are initially trained on either SNLI or MNLI and subsequently tested on the designated target datasets. Δ are absolute differences with BERT _{Base} . 145	145

6.5 Accuracies after performing the Removal of irrelevant features hypothesis with different models 147

6.6 Model parameters and peak memory analysis on different models 147

6.7 Model accuracies analysis without KL loss 148

6.8 Average accuracy over 5 runs with std in parentheses in High resource setting . 152

