Independent Project Winter 2024

Utsav Garg

September 9, 2024

1 Abstract

The report contains the work undertaken during my independent project in the winter semester, primarily centered around training MOS prediction models across various datasets. It extensively explores renowned datasets for MOS prediction, detailing their characteristics and their influence on naturalness. Moreover, it confirms the significance of the newly proposed PRS loss method and extending its application to the SOMOS dataset. Learning more about how loss methods are able to capture audio properties such as vocoding, prosody, and speaker voice. Furthermore, it compares the well-known BVCC dataset with SOMOS, shedding light on their compositional disparities and the respective models learning.

2 Introduction

A long-standing issue in voice conversion, speech synthesis, and speech enhancement systems is the quality quantification of generated speech. Objective metrics, like the Mel-cepstral distance are frequently employed in the speech domain to automatically assess the calibre of transformed speech. However, because these measurements primarily evaluate the distortion of acoustic elements, they are not necessarily connected with human perception. Subjective metrics like the mean opinion score (MOS) and similarity score are able to capture the inherent naturalness and similarity of a speech system; however, because they require a large number of participants to complete listening tests, these assessments are costly, time-consuming, and vulnerable to human bias.

The **Mean Opinion Score (MOS)**, which offers a single numerical result for assessing the perceived quality of media content like VoIP conversations or streaming video, is a commonly used metric in the evaluation of audiovisual and telecommunications quality. Based on factors like naturalness, listeners score samples on a scale of 1 to 5. These ratings are averaged to produce the MOS, which provides an overall evaluation of quality. MOS prediction models are essential for multimedia systems, telecommunications, and audiovisual quality evaluation. They help academics and practitioners in the field understand how encoding methods and transmission problems affect perceived quality. Stakeholders may provide customers with a better watching and listening experience by making educated decisions about display technology, compression algorithms, and content distribution by appropriately forecasting MOS.

The gold standard for assessing synthesised speech is listening tests conducted on human subjects; however, when the number of systems to be evaluated rises, these tests become prohibitively expensive and time-consuming. Furthermore, a number of contextual factors, including the subjectivity of human perception, rater bias, scalability concerns for large-scale assessments, and potential challenges in guaranteeing inter-laboratory consistency among the gathered scores, affect the conventional method of gathering MOS data based on listening tests. These justifications highlight the necessity of an automated, trustworthy MOS prediction process.

3 MOS Datasets

3.1 Blizzard Challenge

In order to better understand and compare research techniques in building corpus-based speech synthesizers on the same data, the Blizzard Challenge has been devised [1]. The basic challenge is to take

the released speech database, build a synthetic voice from the data and synthesize a prescribed set of test sentences. The sentences from each synthesizer will then be evaluated through listening tests. The BC challenges are released every year with the first challenge released in 2005.

3.2 Voice Conversion Challenge

Voice conversion (VC) refers to digital cloning of a person's voice; it can be used to modify audio waveform so that it appear as if spoken by someone else (target) than the original speaker (source). VC is useful in many applications, such as customizing audio book and avatar voices, dubbing, movie industry, teleconferencing, singing voice modification, voice restoration after surgery, and cloning of voices of historical persons. Since VC technology involves identity conversion, it can also be used to protect the privacy of the individual in social media and sensitive interviews, for instance. For the same reason, VC also enables spoofing (fooling) voice biometric systems and has therefore potential security implications. The VCC2020 challenge [8], similar to the two earlier editions of the challenge, does not focus on any particular application but aims at improving the core VC technology itself using common data, metrics and baseline systems provided by the organizers.

3.3 VoiceMOS Challenge

Human listening tests are the gold standard for evaluating synthesized speech. Objective measures of speech quality have low correlation with human ratings, and the generalization abilities of current data-driven quality prediction systems suffer significantly from domain mismatch. The VoiceMOS Challenge [3] aims to encourage research in the area of automatic prediction of Mean Opinion Scores (MOS) for synthesized speech. This challenge has two tracks:

- Main track: We recently collected a large-scale dataset of MOS ratings for a large variety of text-to-speech and voice conversion systems spanning many years, and this challenge releases this data to the public for the first time as the main track dataset.
- Out-of-domain track: The data for this track comes from a different listening test from the main track. The purpose of this track is to study the generalization ability of proposed MOS prediction models to a different listening test context. A smaller amount of labeled data is made available to participants, and unlabeled audio samples from the same listening test are made available as well, to encourage exploration of unsupervised and semi-supervised approaches.

3.4 DNS Challenge

The Deep Noise Suppression challenge [2] encompasses a comprehensive set of tasks including noise suppression, de-reverberation, and interference suppression for both headset and non-headset scenarios. Divided into two tracks, it focuses on enhancing speech quality for various audio devices such as headphones, speakerphones, and built-in microphones in different devices. Evaluation criteria are based on the ITU-T P.835 subjective test framework, modified to ensure reliability in scenarios involving interfering talkers. Metrics su4r5ch as Speech Quality (SIG), Background Noise Quality (BAK), Overall Audio Quality (OVRL), and Word Accuracy (WAcc) are utilized to measure model performance.

3.5 ASVSpoof Challenge

The automatic speaker verification spoofing and countermeasures (ASVspoof) challenge series is a community-led initiative which aims to promote the consideration of spoofing and the development of countermeasures. We have data for ASVSpoof2019. Will have to check whether the ASVSpoof2021 [12] listening test has MOS values. Its main evaluation metric involves min t-DCF, ERR.

3.6 NISQA Corpus

The NISQA Corpus [5] includes more than 14,000 speech samples with simulated (e.g. codecs, packet-loss, background noise) and live (e.g. mobile phone, Zoom, Skype, WhatsApp) conditions. Each file is labeled with subjective ratings of the overall quality and the quality dimensions Noisiness, Coloration, Discontinuity, and Loudness. In total, it contains more than 97,000 human ratings for each of the dimensions and the overall MOS.

3.7 LibriTTS

The new corpus inherits desired properties of the LibriSpeech corpus while addressing a number of issues which make LibriSpeech less than ideal for text-to-speech work. The released corpus consists of 585 hours of speech data at 24kHz sampling rate from 2,456 speakers and the corresponding texts. Experimental results show that neural end-to-end TTS models trained from the LibriTTS [13] corpus achieved above 4.0 in mean opinion scores in naturalness in five out of six evaluation speakers (Released by Google AI - Interspeech 2019) LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech, prepared by Vassil Panayotov with the assistance of Daniel Povey. The data is derived from read audiobooks from the LibriVox project, and has been carefully segmented and aligned.

Dataset Name	Language	Systems	Utterances
BC2023 (Spoke) - a	French	21	882
BC2023 (Hub) - b	French	17	578
SVCC23 (In-dom) - c	English	25	2000
SVCC23 (In-dom) - d	English	24	1920
Noisy & enhanced - e	Chinese	97	1940
VoiceMOS2023	a+b+c+d+e	183	7320
NISQA	English, German	-	14000

4 SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis

SOMOS [4] is the first resource consisting of speech entirely produced by neural TTS models and their MOS ratings, adequately sized so as to train deep learning based system. It is a single english speaker dataset trained on LJ Speech Dataset. It employs a single LPCNet Vocoder and Tacotron based TTS systems for the production of speech signals. The role of **neural vocoders** in TTS is to synthesize human-like speech signals from Acoustic Features such as the spectrogram in the mel scale and to produce audio waveform. **Tacotron2** is an end-to-end generative text-to-speech model that takes a character sequence as input and outputs the corresponding Mel spectrogram. The backbone of Tacotron is a seq2seq model with attention. The SOMOS dataset focuses on **voice prosody**. It includes features such as intonation, stress, rhythm, and pauses, which contribute to the overall expressiveness and naturalness of speech"

4.1 SOMOS Sentences

The corpus has a variety of linguistic contents for synthesis, so the models that would be trained on the generated speech corpus would be able to generalize over it. The corpus comprises 2,000 sentences, out of which, 100 were randomly selected from the LJ Speech script and excluded from the LJ Speech training dataset. The remainder of the corpus is composed in such a way to ensure domain, length and phoneme coverage. English sentences from the Blizzard Challenges of years 2007-2016 were selected, mainly from the conversational, news, reportorial and novel domains, as well as 100 semantically unpredictable sentences (SUS). Additionally, Wikipedia and "general" public domain sentences from the web, which are common for conversational agents, have been included. The length of the corpus' sentences ranges from 3 to 38 words, with a median of 10 words. In our speech dataset, each system utters 100 sentences and each sentence is uttered by 10 distinct systems.

4.2 SOMOS Score Collection

All subjective evaluations were conducted online with crowdsourced naive listeners via Amazon Mechanical Turk (AMT) with measures to maintain quality of ratings. Although the expert's scores average lower than crowdsourced scores, a large system level positive relation is observed with all

locales' results, using both the Pearson (PCC) and the Spearman rank correlation coefficient (SRCC). This suggests that the test design has led to consistent answers that match the expert judgements. The locale with strongest correlation to expert ratings is GB (SRCC at r = 0.88, PCC at r = 0.84), while the weakest correlation is observed in US (SRCC at r = 0.85, PCC at r = 0.81). Interestingly, the correlation to the expert gets stronger for the entire dataset (SRCC at r = 0.90) suggesting the importance of combining ratings of several workers, as also showcased in [38, 39].

5 Difference in SOMOS and BVCC

The state of art results on BVCC using the proposed PRS [11] loss method encourages to further proof test its application on newer datasets. The SOMOS data is different from the BVCC dataset in multiple aspects and allows us to gain insights of what features of an audio resembles more to human and natural sounds. Major differences in both the datasets are listed below:

SOMOS	**BVCC**
Single speaker	Multiple speakers
Same vocoder	Multiple vocoders
Learns which prosodic variations are more human-like	Learns which speaker, vocoder is better suited for human-like voice generation
20,000 utterances	5,000 utterances
Huge amount of training data allows the model to learn the prosody effectively	Lesser amount of training data with combination of different MOS Challenges therefore does'nt learn about prosody.

Different speaking style and vocoding ability affects the understanding and the mos scores for an audio, therefore the BVCC dataset focuses on the generalization ability of SSL models for MOS prediction rather than understanding the prosody which correlates with natural voices.

6 Experiments and Results

Metric	Pretrained L1	Pretrained PRS	Trained PRS
MSE	0.222518	3.163443	6.509897
LCC	0.873842	0.876994	0.874864
SRCC	0.875749	0.880515	0.876815
KTAU	0.699872	0.707091	0.701459

Figure 1: Utterance level results on stage 1 training using BVCC dataset

ASV 2019				BC 2019				COM2018 (XW2018)			
MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
2.750	0.470	0.499	0.357	19.51	0.637	0.686	0.500	6.34	0.515	0.490	0.350

Figure 2: Zero Shot Results of Utterance level trained on BVCC dataset

Labelled Samples	UnLabelled Samples	-	5	10	20
Semi-Supervised Setting					
0	136	0.778			
0	676	0.778			
Semi-Supervised Setting + BApMOS Selection					
0	136		0.767	0.797	
0	676				0.795

Figure 3: Stage 2 BapMOS training results on BVCC dataset

	Utterance Level			System Level		
Model	MSE	LCC	SRCC	MSE	LCC	SRCC
PRS(full)	7.6844	.5760	.6447	7.4326	.7846	.8528
PRS(clean)	6.191	0.6128	0.6520	5.7298	0.8310	.8750
L1 (full)	0.3352	0.6631	0.6548	0.1911	0.8760	0.8883
L1 (clean)	0.2284	0.6605	0.6513	0.051	0.8876	0.8947

Figure 4: Stage 1 training results on SOMOS dataset

The SOMOS paper demonstrated a very better system level scores for SSL-MOS than our model which should not be the case. On detailed reading of the SOMOS paper it gave that the results on the paper are based on train-test split of 85-15% whereas the dataset released has train-dev-test set with ratio of 70-15-15. To further test the hypothesis, I created python script which randomly makes the split of 85-15% according to the constraints mentioned in the paper. On training and testing according to this split, we got comparable and even better results in some cases.

Dataset Trained On	Validation/Test Set	Results							
		Utterances				System			
		MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
SOMOS-clean	Custom Created #1	1.00322	0.547518	0.670529	0.48732	0.328243	0.923375	0.942815	0.798387
SOMOS-clean	Custom Created #2	1.241688	0.550571	0.681663	0.493565	0.454412	0.91582	0.946848	0.810484
SOMOS-clean	Custom Created #3	1.086165	0.545322	0.682889	0.496699	0.371662	0.910955	0.884897	0.717742
SOMOS-clean	Custom Created #4	1.479498	0.604292	0.713309	0.521898	0.581318	0.966648	0.957478	0.842742
SOMOS-clean	Custom Created #5	1.499261	0.568213	0.71961	0.526342	0.526895	0.946439	0.938416	0.802419
SOMOS-clean	Custom Created #Avg	1.2619664	0.5631832	0.6936	0.5051648	0.452506	0.9326474	0.9340908	0.7943548
SOMOS-full	Custom Created #1	7.681321	0.585573	0.654975	0.468579	7.482298	0.940464	0.939883	0.794355
SOMOS-full	Custom Created #2	7.779999	0.569297	0.629638	0.449416	7.602838	0.972955	0.955279	0.8508
SOMOS-full	Custom Created #3	7.919365	0.580818	0.650799	0.469127	7.742495	0.835778	0.932918	0.78629
SOMOS-full	Custom Created #4	7.654049	0.565973	0.628393	0.449022	7.442042	0.920949	0.91349	0.745968
SOMOS-full	Custom Created #5	7.897559	0.578967	0.670996	0.484986	7.733	0.871231	0.955279	0.818548
SOMOS-full	Custom Created #Avg	7.7864586	0.5761256	0.6469602	0.464226	7.6005346	0.9082754	0.9393698	0.7991922

Figure 5: Custom Split created and their utterance and system level results

Model	Dataset Trained On	Validation/Test Set	MSE	LCC	SRCC	KTAU
PRS	SOMOS Clean	Testset (ASV2019)	1.944184	0.223723	0.26571	0.187276
PRS	SOMOS Full	Testset (ASV2019)	4.245646	0.249329	0.261391	0.184822
L1	BVCC	Testset (ASV2019)	1.498	0.47	0.491	0.352
PRS	BVCC	Testset (ASV2019)	2.75	0.47	0.499	0.357
PRS	SOMOS Clean	Testset (BC2019)	47.463307	0.286755	0.288604	0.197182
PRS	SOMOS Full	Testset (BC2019)	12.897606	0.358483	0.272182	0.184336
L1	BVCC	Testset (BC2019)	3.672	0.553	0.559	0.409
PRS	BVCC	Testset (BC2019)	19.51	0.637	0.686	0.5
PRS	SOMOS Clean	Testset (COM2018)	27.098699	0.112481	0.129261	0.088925
PRS	SOMOS Full	Testset (COM2018)	6.92627	0.088478	0.081331	0.055442
L1	BVCC	Testset (COM2018)	1.2	0.476	0.423	0.297
PRS	BVCC	Testset (COM2018)	6.34	0.515	0.49	0.35

Figure 6: Comparison of results on out of domain zeroshot testing using pretrained bvcc and somos checkpoint.

The BVCC trained model works very better than the SOMOS trained dataset. First because the BVCC dataset is formed for generic prediction of MOS scores and contains samples from various different challenges. Secondly both ASV and BC challenge focuses on human spoken tasks and has nothing to do with prosodic variation learning, therefore SOMOS trained model is expected to perform lesser.

	Semi Value	Learning Rate	MSE	LCC	SRCC	KTAU
EPOCH 0	-	-	3.486296	0.545425	0.581703	0.416288
EPOCH 44	50	1.1	3.487839	0.545486	0.581909	0.416385
EPOCH 15	50	2.1	3.223187	0.580992	0.62244	0.448909
EPOCH 0	15	2	3.486211	0.545375	0.58154	0.416119
EPOCH 57	10	2.1	6.643569	0.583789	0.600654	0.429441

Doing out of domain testing for the new checkpoint ->

Test Set	MSE	LCC	SRCC	KTAU
ASV2019	3.329798	0.241492	0.251047	0.176549
BC2019	13.357091	0.502772	0.392899	0.268483
COM2018	6.299593	0.295473	0.306328	0.212638

Figure 7: Stage 2 training on BVCC using Stage 1 SOMOS PRS CLEAN checkpoint

		Val Set Results (SOMOS L1 CLEAN)				
Semi Value	Learning Rate	Epoch No	MSE	LCC	SRCC	KTAU
-	-	EPOCH 0	1.018746	0.533994	0.561867	0.393904
50	1.1	EPOCH 159	1.281992	0.542472	0.570927	0.403012
50	2.1	EPOCH 11	1.202187	0.55908	0.568907	0.400528
15	2	EPOCH 12	0.814745	0.631871	0.628136	0.45205
10	2.1	EPOCH 104	1.270173	0.541504	0.56732	0.399103

Doing out of domain testing for the new checkpoint ->

Test Set	MSE	LCC	SRCC	KTAU
ASV2019	2.132653	**0.252681**	**0.256168**	**0.180561**
BC2019	1.960439	**0.519386**	**0.463134**	**0.321476**
COM2018	0.848261	0.153994	0.171887	0.118662

Figure 8: Stage 2 training on BVCC using Stage 1 SOMOS L1 CLEAN checkpoint

The stage 2 training on BVCC using the stage 1 SOMOS checkpoint showed the increase in the correlation for BC challenge and COM challenge. Giving a positive and faster method to finetune a model with little data for application purposes. Although both the checkpoints show similar results. The L1 checkpoint works a little better for stage 2 finetuning but almost comparable.

Model	Test Set	MSE	LCC	SRCC	KTAU
SOMOS_L1_BC2019	BVCC	1.601979	**0.542308**	**0.595621**	**0.421428**
(SOMOS FULL L1)	ASV2019	3.286204	0.209908	0.221649	0.155655
	BC2019	0.812708	0.499392	0.500439	0.351894
	COM2018	0.851577	0.194959	0.202422	0.139476
SOMOS_PRS_BC2019	BVCC	10.496547	0.433704	0.393272	0.27456
(SOMOS FULL PRS)	ASV2019	6.650196	0.096353	0.104069	0.073006
	BC2019	10.563346	**0.54989**	**0.571853**	**0.412473**
	COM2018	11.304626	0.240651	0.162838	0.114041

Figure 9: Stage 2 training on BC2019 using Stage 1 SOMOS L1 CLEAN checkpoint

			Utterance Level			
Test Set	**Checkpoint**		**MSE**	**LCC**	**SRCC**	**KTAU**
SOMOS Clean	BVCC L1		**1.247884**	**0.422227**	**0.415891**	**0.285458**
SOMOS Clean	BVCC PRS		10.396044	0.376874	0.359501	0.246732
SOMOS Full	BVCC L1		**1.561842**	**0.411519**	**0.402781**	**0.276838**
SOMOS Full	BVCC PRS		11.592244	0.375506	0.352072	0.241733

Figure 10: Zero shot result of BVCC trained model on SOMOS dataset

The above table confirms our little hypothesis that for the BVCC dataset, the pretrained L1 model is better in understanding the prosodic variations of an audio in comparsion to pretrained PRS.

		Utterances		
Train Set	**Factor**	**MSE**	**LCC**	**SRCC**
SOMOS Clean	**0.5**	9.102298	**0.664955**	**0.663299**
SOMOS Full	**0.5**	10.047408	0.64627	0.635208
SOMOS Clean	**0.1**	6.191	0.6128	0.652
SOMOS Full	**0.1**	7.6844	0.576	0.6447
L1 Clean	-	0.2284	0.6605	0.6513
L1 Full	-	0.3352	**0.6631**	**0.6548**

Figure 11: Training SOMOS dataset with factor (lambda c) = 0.5

7 Conclusion and Learning

- Learned about various datasets and challenges, from voice conversion to text to speech to noise compression to automatic spoof detection in speech.
- The SOMOS dataset was really insightful for detailed and deep understanding of how TTS system works. Why are vocoders used, what all are present in speech and how do you represent all that through text with the use of TOBI labels.
- Learned about HOW RESEARCH IS DONE? How we compare two datasets learn whats different in them and accordingly understand how architectural change learns about specific properties of speech.
- Major part of my IP was to train model and understand, debug the code. The PRS code on the github helped me learn how are deep learning codes written from superficial to deep level, be it data loader creation to batch size importance, or the way the evaluation metrics are reported for each epoch to understand how the model is learning and make changes effectively.
- In terms of work, I learned how to inference pretrained models. (Figure 1)
- I trained the SOMOS model from scratch using the PRS and could report the great performance of the model of SRCC equivalent to 0.65 with lesser training data that is used in the offical paper where their SSL-MOS model could just get around 0.4. Showed that PRS is able to effectively learn about the prosodic variations in speech. (Figure 4)

- Also confirmed the effectiveness of BVCC for generalized understanding of speech for MOS prediction in comparison to SOMOS which focuses only on prosody. (Figure 6)
- For BC2019 and COM2018, the stage 2 finetuning of BVCC on SOMOS pretrained checkpoint improved the SRCC values significantly, depicting their closness to BVCC than SOMOS in terms of training data. (Figure 7,8)
- The zeroshot test on BVCC using SOMOS PRS and L1 checkpoint, shows better performance of L1 checkpoint. (Figure 10).

Few learnings are mentioned in the bottom of Related Work

8 Future Work

- Understand why for few datasets L1 trained checkpoint is better at fine tuning in comparison to PRS.
- Is L1 better in understanding prosodic variations, how are these variations represented in audio?
- Improve the stage2 BapMOS training using ensemble learning and other techniques.

9 Related Work - Literature Review

9.1 Uncertainty as a Predictor: Leveraging Self-Supervised Learning for Zero-Shot MOS Prediction [6]

9.1.1 Statement of Purpose and Objectives

Predicting audio quality in voice synthesis and conversion systems is a challenging task while using traditional method like Mean Opinion Scores (MOS). This paper explores the surprising ability of out-of-the-box pre-trained SSL models on zero-shot MOS prediction. They hypothesize that:

- Uncertainty estimates can be derived from the outputs of SSL models such as wav2vec, which can be used as proxies to MOS scores.
- This is based on the fact that a high uncertainty values around the contents of an audio sequence must correspond to low audio quality.

The main contribution of the paper is to create these uncertainty measures, which will reflect the quality of the tokens uttered in an audio sample. It also shows how these measures correlates with intelligibility measures and MOS scores.The scope of the paper is to provide insights into why SSL models seem to work well for MOS prediction and provides a totally zero-shot baseline for low-resource MOS prediction settings.

9.1.2 Working and Obtaining the UMs

- Pass the audio through SSL model to get w,q sized matrix representation where w is time windows and q is the size of the logits returned. (Contrastive predictive logits or token utterance probabilities.)
- These logits are softmaxed and probabilities of the logits are obtained.
- These probability scores are used to calculate the different uncertainty measures such as entropy, max, mean and standard deviation.

The maximum logit is a good proxy as one would expect the entropy of a categorical distribution to be low if one of the logits is very high (which would represent high confidence in the emission of a token. In the cases where a clear probabilistic interpretation of the output layer is not available, such as our experiments where a wav2vec2 model is available but not with a suitable ASR head or quantizer, we treat the encoder output as logits of a Categorical distribution, as these are typically fed to a quantizer. These uncertainty measures are computed for each audio sample and the SRC coefficients between these uncertainties and associated MOS values are calculated.

Table 1. Experimental results of our experimental setup using spoken English data

model	type	mean	max	sd	entropy
wav2vec	Large	**-69.1**	**68.2**	64.0	**-69.9**
	VQ	-60.1	58.4	**68.3**	-69.0
wav2vec2 ASR	base 10m	6.5	23.4	26.8	-23.0
	base 100h	-23.8	45.5	46.3	**-39.2**
	base 960h	**-45.7**	**50.6**	**52.4**	-37.3
	large 960h	-9.9	39.9	40.3	-25.7
	lv60k 960h	32.4	-1.6	45.8	-37.4
	voxpop. en	-39.0	32.1	42.7	-43.6

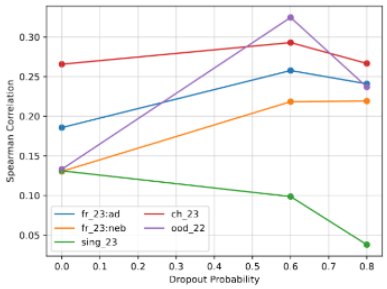

Figure 12: The uncertainty measures results

9.1.3 Experiment Results

The wav2vec section of Table 1 shows that all our UMs correlate strongly with MOS scores, achieving SRCC values of about 70%. The baseline model of [1] achieved an SRCC of about 92% on this task, so while our predictors are not state-of-the-art, they are nonetheless strong predictors given that no task-specific model training was done. Also the UMs values are signs consistent as well. As high entropy suggests greater uncertainty, while high maximum shows confidence and thus lower uncertainty. Similarly for standard deviation, less deviation shows greater confidence.

9.1.4 Conclusion and Future Work

This paper has demonstrated the potential of using uncertainty measures from SSL models for predicting MOS rating. wav2vec is notably the most performant for the tasks and shows that a moderately strong predictor of audio quality can be achieved in no-resource settings. Future exploration could extend these insights to other domains where understanding out-of-domain data is critical, such as dimensionality reduction methods, which are known to have probabilistic interpretations reminiscent of self-supervised methods

9.2 Partial Rank Similarity Minimization Method for Quality MOS Prediction of Unseen Speech Synthesis Systems in Zero-Shot and Semi-supervised Setting [11]

9.2.1 Introduction

Speech synthesis systems represent an advanced technology for generating artificial speech. These systems have evolved to address various challenges, including the ability to generate speech for languages or tasks not explicitly included in the training data, known as the zero-shot setting. In this scenario, the system is trained to generalize and produce speech for unseen languages or tasks, demonstrating its adaptability across diverse linguistic contexts. Similarly, in semi-supervised settings, where systems are trained on a mix of labeled and unlabeled data, they exhibit enhanced generalization and adaptability across different speech domains.

9.2.2 Other MOS Predictor

Evaluation of speech synthesis systems often relies on the Mean Opinion Score (MOS), a metric reflecting the perceived quality of generated speech by human listeners. MOS provides a subjective measure based on human judgment, offering insights beyond technical metrics. MOSNet, a neural network model, correlates highly with human MOS ratings, with notable improvements observed when employing transformer architectures over traditional LSTM or CNN models.

9.2.3 Proposed PRS Loss Function

To further enhance the evaluation and performance of MOS prediction models, novel loss functions have been proposed. The Partial Rank Similarity (PRS) method outperforms traditional L1 loss, particularly in zero-shot and semi-supervised settings. PRS focuses on preserving the relative ranking of

speech samples rather than solely considering absolute MOS values, thus providing a more meaningful evaluation criterion. Motivated by the significance of relative ranking, PRS incorporates a partial rank order within a mini-batch, emphasizing the correct order of samples in training. The objective function minimizes total losses with respect to the training data, penalizing misclassifications of sample order to encourage learning of correct rank orders. In practical implementation, limitations such as GPU memory constraints are addressed by variants such as Extended PRS (E-PRS), which stores outputs of previous batches for comparison with current batch samples. Moreover, the proposal to set lambda based on differences in PR ensures a balanced distribution of pseudo MOS values, enhancing the weighting of MOS values distant from the ground truth.

$$\mathcal{L}_{PRS} = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda * |PR_{ij}(\hat{\boldsymbol{Y}}) - PR_{ij}(\boldsymbol{Y})|^p \right)^{1/p} \tag{2}$$

Where:

$$\hat{\boldsymbol{Y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n), \tag{3}$$

$$\hat{y}_i = \Psi(x_i), \tag{4}$$

$$\lambda = \begin{cases} 1 & \text{if } \{PR_{ij}(\hat{\boldsymbol{Y}}) \cdot PR_{ij}(\boldsymbol{Y})\} \leq 0, \\ \lambda_c <= 1 & \text{otherwise,} \end{cases} \tag{5}$$

Figure 13: The novel loss function - PRS

9.2.4 BapMOS Algorithm

The proposed method addresses the challenge of selecting accurate pseudo Mean Opinion Score (MOS) values for audio samples lacking labeled data in semi-supervised learning settings. Initially, the model is trained using labeled samples through supervised learning. Then, pseudo MOS values are estimated for unlabeled samples using the trained model. These pseudo MOS values are merged with labeled samples for iterative supervised learning. However, selecting all unlabeled samples can destabilize subsequent training phases due to inaccurate pseudo MOS values. To mitigate this, the proposed Balanced pseudo MOS (BApMOS) selection algorithm balances the distribution of pseudo MOS values. It constructs a histogram with bins and randomly samples the minimum count of pseudo MOS values from each bin to ensure a balanced distribution. This approach enhances the reliability of pseudo MOS values used in iterative training, contributing to improved model performance.

9.2.5 Conclusion

In summary, advancements in speech synthesis systems and MOS prediction models underscore the importance of not only technical performance metrics but also subjective human judgment in assessing the quality of synthesized speech. The development of novel loss functions and evaluation methodologies reflects ongoing efforts to improve the robustness and applicability of these technologies across diverse linguistic and contextual domains.

9.3 Tacotron: Towards End-to-End Speech Synthesis [10]

Tacotron, an end-to-end generative text-to-speech model that synthesizes speech directly from characters. Given pairs, the model can be trained completely from scratch. Tacotron achieves a 3.82 subjective 5-scale mean opinion score on US English, outperforming a production parametric system in terms of naturalness and it is faster.

$$Text \rightarrow MODEL \rightarrow \text{Raw Spectrograms}$$

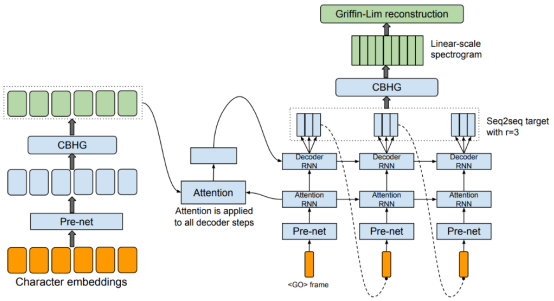

Figure 14: The tacotron architecture

9.3.1 Other TTS Model

WaveNet: TTS but slow due to autoregressive nature. Not end-to-end DeepVoice: Replace every component of TTS with a neural network however each network is trained separately, the multi-stage model leads to compounded component's errors. Char2Wav: Their seq2seq and SampleRNN needs separate pre-trained, does not allow it from scratch.

9.3.2 Architecture

- **Encoder**: The character sequence is fed into the model by using the one-hot encoding to get a continuous embedding. This embedding is fed into the CBHG module which transforms it to encoder output representation.
- **CBHG Module**: The module extracts robust sequential representations of text. It does this by:
  - **1-D Convolutional Filters**: The input sequence undergoes convolution with multiple sets of 1-D filters. Each set captures different widths of contextual information, similar to modeling unigrams, bigrams, up to K-grams.
  - **Max Pooling**: The convolution outputs are stacked and max-pooled along the time dimension to enhance local invariances while preserving the original time resolution.
  - **Residual Connections**: The processed sequence is augmented with the original input sequence using residual connections, aiding in gradient flow during training.
  - **Highway Networks**: The convolution outputs are fed into a multi-layer highway network, enabling the extraction of high-level features while mitigating the vanishing gradient problem.
  - **Bidirectional GRU RNN**: A bidirectional GRU RNN is stacked on top to capture sequential features from both forward and backward contexts, providing comprehensive context understanding.
- **Decoder**: How the decoder works:
  - **Content-Based Tanh Attention Decoder**: Utilizes a recurrent layer to produce the attention query at each decoder time step, combining it with the context vector and attention RNN cell output as input to the decoder RNNs.
  - **Decoder RNNs**: Employs a stack of GRUs with vertical residual connections, which accelerate convergence.
  - **Target Representation**: Instead of predicting raw spectrograms, uses a compressed target representation, such as an 80-band mel-scale spectrogram, to facilitate learning alignment between speech signal and text.
  - **Prediction Strategy**: Predicts multiple non-overlapping output frames at each decoder step, which speeds up convergence by allowing attention to move forward early in training. During inference, the last frame of the predictions is fed as input to the next decoder step.

- **Training Procedure**: During training, every r-th ground truth frame is fed to the decoder, with dropout in the pre-net to aid generalization by providing noise to resolve multiple modalities in the output distribution.

- **Post Processing Net**: The post-processing net's task is to convert the seq2seq target to a target that can be synthesized into waveforms. They use Griffin-Lim as the synthesizer, the post-processing net learns to predict spectral magnitude sampled on a linear-frequency scale. The network chosen is again CBHG as it allows both forward and backward information to correct the prediction error for each individual frame. The Griffin-Lim algorithm (Griffin and Lim, 1984) to synthesize waveform from the predicted spectrogram.

9.3.3 Results

Tacotron, an integrated end-to-end generative TTS model that takes a character sequence as input and outputs the corresponding spectrogram. With a very simple waveform synthesis module, it achieves a 3.82 MOS score on US English, outperforming a production parametric system in terms of naturalness. Tacotron is frame-based, so the inference is substantially faster than sample-level autoregressive methods. Unlike previous work, Tacotron does not need handengineered linguistic features or complex components such as an HMM aligner. It can be trained from scratch with random initialization.

9.4 LPCNet: Improving Neural Speech Synthesis Through Linear Prediction [9]

Neural speech synthesis models have recently demonstrated the ability to synthesize high-quality speech for text-to-speech and compression applications. However, these new models often require powerful GPUs to achieve real-time operation, limiting their deployment on lower-power devices such as embedded systems and mobile phones. To address this issue, researchers have proposed LPCNet, a WaveRNN variant that combines linear prediction with recurrent neural networks to significantly improve the efficiency of speech synthesis. The key idea behind LPCNet is to leverage linear prediction, a well-established technique in speech coding, to reduce the complexity of the neural network required for speech synthesis.

9.4.1 What is a Vocoder?

The role of neural vocoders in TTS is to generate/synthesize natural/human-like speech signals from Acoustic Features (AF). The input of neural vocoders is usually a certain type of acoustic features, such as the spectrogram in the mel scale or in the bark scale, etc. Acoustic features are supposed to be a salient yet compact representation of the speech signal. A rich collection of studies show that human ears are more sensitive in lower frequencies than higher frequencies. Consequently, a spectrogram in the mel scale is deemed to be more aligned with human auditory perception. The most important aspect of the vocoder is its ability to make a synthesized sound mimic a real world sound. Using fourier inversion it can convert digital signals into analogous signals.

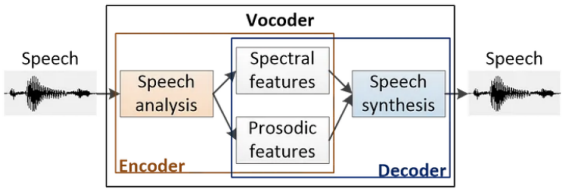

Figure 15: The working of a vocoder

To get MFCC or BFCC with cepstral analysis which are usually the input (or extracted feature) of the neural synthesis systems. The inputs are acoustic features and the output is speech waveform.

9.4.2 Architecture

- Leveraging Linear Prediction (LP): LPCNet incorporates Linear Prediction, a signal processing method, to represent speech as a filter and its excitation source. This simplifies the task for the deep neural network (DNN) model within LPCNet.
- WaveRNN Variant: LPCNet builds upon WaveRNN, another neural vocoder architecture. It inherits properties from WaveRNN like sparse matrices to improve efficiency.
- Frame-rate and Sample-rate Networks: LPCNet utilizes two separate networks: Frame-rate network: predicts a conditioning feature based on input features. Sample-rate network: predicts the probability distribution for the excitation signal using the conditioning feature from the frame-rate network along with other inputs.

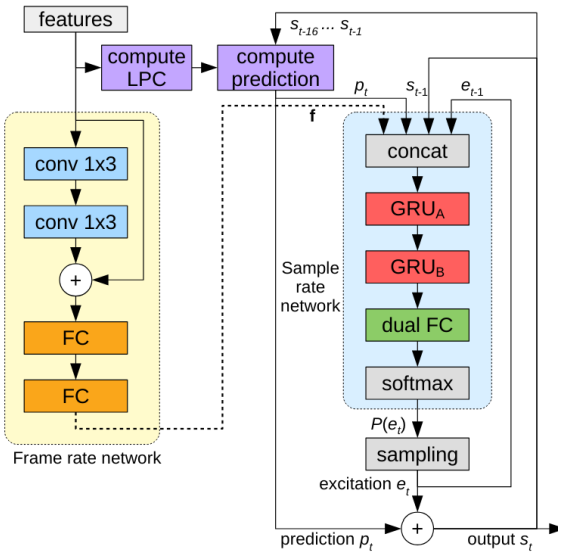

Figure 16: LPCNet Architecture

9.4.3 Results

The authors of the LPCNet paper demonstrate that high-quality speech synthesis is achievable with a complexity under 3 GFLOPS, making it easier to deploy neural synthesis applications on lower-power devices. They also show that LPCNet can outperform other efficient neural speech synthesis models, such as WaveRNN and LPCNet-based models, in terms of both quality and computational efficiency. They have demonstrated an improvement in synthesis quality while operating 2.5x faster, resulting in an open-source LPCNet algorithm that can perform real-time neural synthesis on most existing phones and even some embedded devices.

9.5 System Variations in SOMOS

9.5.1 How are the 200 systems created?

Basically Tacotron based systems are used for creating the different systems. The basic vanilla model used by SOMOS is [3,4]. Which in detail tells about the procedure for speech generation from text. It also uses:

- **Through Linguistic Annotation**: One is enriching the text with noun chunks boundaries and the second is embedding part of speech (POS) information. The models are **generated implicitly**, indicating that these changes are not explicitly programmed but rather inferred from linguistic annotations, which are added labels providing information about the linguistic structure of the text.

- **TOBI (Tones and Break Indices)** [7]: is a system for transcribing and annotating the intonation patterns and other aspects of the prosody of English utterances inputted parallel to the input phoneme sequence. It adds these prosodic variations to the text by tones which represent the analysis of tonal events in terms of H and L. Tonal events include pitch accents, phrase accents, and boundary tones. Break indices mark the prosodic grouping of the words in an utterance and show the degree of disjuncture between adjacent words on a scale from 0 to 4. ToBI dataset includes not only text-speech pairs but also ToBI labels.

SOMOS uses the first end-to-end English speech synthesis system combined with ToBI representation. The ToBI prediction module uses a pretraining language model to predict ToBI labels which as input features consist of phoneme, lexical stress, break index, pitch accent, phrase accent, and boundary tone, and the Tacotron model is used for prosody modeling.

9.5.2 Working of the Architecture

In the training phase, the ToBI labels are used as the target of the frontend network as well as one of the input to the acoustic model. In the inference phase, the frontend converts the text sequence to ToBI tags, and then the phoneme sequence and the predicted ToBI tags are fed to the backend TTS model to generate speech.

9.5.3 ToBI Frontend

ToBI prediction frontend aims to automatically predict the appropriate prosody from the text, including pause, stress, and intonation. Using the pre-trained model ELECTRA is conducive to extracting the grammatical and semantic information in the text, which is helpful for the prediction of ToBI. Compared with the original Tacotron, we also use ToBI labels as a part of the input features to the encoder, considering that encoder should be able to capture text-related prosodic information. Overall, the input features consist of phoneme, lexical stress, break index, pitch accent, phrase accent, and boundary tone. By introducing ToBI features into training, we can control the system to synthesize speech with different prosodies by giving different ToBI labels during inference.

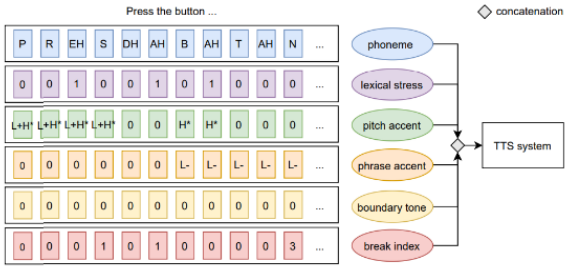

Figure 17: Addition of ToBI labels to a TTS system

9.5.4 Additional Definitions

- **Syntactic cues** could include information about sentence structure, grammatical relationships between words, or other syntactic features derived from linguistic analysis.
- **Noun chunks boundaries** Noun chunk boundaries in a text refer to the delineation of phrases or segments that contain a noun and its related modifiers. Essentially, noun chunks are contiguous sequences of words that function together to convey a single concept or idea, with a noun serving as the head of the chunk.
- **Pitch accents** mark the stressed syllable of specific words that carry the most information in a sentence. Pitch accents include H* (high accent), L* (low accent), L*+H (a syllable which starts with a low accent and then rises) and L+H* (again low-high on one syllable, but with the second part accented).

- **Boundary tones** describe the pitch trend at each full intonation phrase boundary. The default is H% (high tone) and L% (low tone).
- **Phrase accents** describe the pitch movement between the ultimate pitch accent and the boundary tone. The default is H- and L-.
- **Break indices** represent the degree of disjuncture between adjacent words, examples of which are 4 for a full intonational phrase break, marked with L% or H%, at the end of a phrase or sentence, 3 for an intermediate phrase break, marked with H- or L-, and 1 for most phrase-internal word boundary.

  Phrase accents and boundary tones together form phrasal tones which affect intonations of utterance, and pitch accents and break indices affect stress and pause of utterance

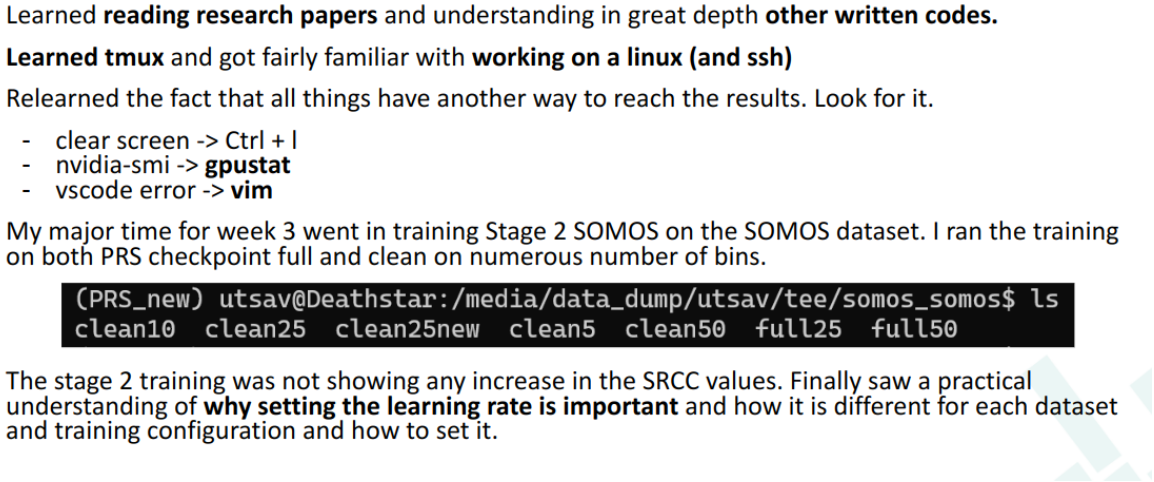

Figure 18: Few other things that I learned during my IP

References

[1] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. Generalization ability of mos prediction networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8442–8446. IEEE, 2022.

[2] Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Alex Ju, Mehdi Zohourian, Min Tang, Hannes Gamper, et al. Icassp 2023 deep speech enhancement challenge. *arXiv preprint arXiv:2303.11510*, 2023.

[3] Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. The voicemos challenge 2022. *arXiv preprint arXiv:2203.11389*, 2022.

[4] Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis. Somos: The samsung open mos dataset for the evaluation of neural text-to-speech synthesis. *arXiv preprint arXiv:2204.03040*, 2022.

[5] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*, 2021.

[6] Aditya Ravuri, Erica Cooper, and Junichi Yamagishi. Uncertainty as a predictor: Leveraging self-supervised learning for zero-shot mos prediction. *arXiv preprint arXiv:2312.15616*, 2023.

[7] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. Tobi: A standard for labeling english prosody. 10 1992.

[8] Patrick Lumban Tobing, Yi-Chiao Wu, and Tomoki Toda. Baseline system of voice conversion challenge 2020 with cyclic variational autoencoder and parallel wavegan. *arXiv preprint arXiv:2010.04429*, 2020.

[9] Jean-Marc Valin and Jan Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5891–5895. IEEE, 2019.

[10] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[11] Hemant Yadav, Erica Cooper, Junichi Yamagishi, Sunayana Sitaram, and Rajiv Ratn Shah. Partial rank similarity minimization method for quality mos prediction of unseen speech synthesis systems in zero-shot and semi-supervised setting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE, 2023.

[12] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[13] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.