SQUID 🌸: MEASURING SPEECH NATURALNESS IN MANY LANGUAGES

Thibault Sellam, Ankur Bapna, Joshua Camp, Diana Mackinnon, Ankur P. Parikh, Jason Riesa

Google

ABSTRACT

Much of text-to-speech research relies on human evaluation. This incurs heavy costs and slows down the development process, especially in multilingual applications where recruiting and polling annotators can take weeks. We introduce SQuId (Speech Quality Identification), a multilingual naturalness prediction model trained on over a million ratings and tested in 65 locales—the largest effort of this type to date. The main insight is that training one model on many locales consistently surpasses mono-locale baselines. We show that the model outperforms a competitive baseline based on w2v-BERT and VoiceMOS by 50.0%. We then demonstrate the effectiveness of cross-locale transfer during fine-tuning and highlight its effect on zero-shot locales, for which there is no fine-tuning data. We highlight the role of non-linguistic effects such as sound artifacts in cross-locale transfer. Finally, we present the effect of model size and pre-training diversity with ablation experiments.

1. INTRODUCTION

Evaluation is a major bottleneck for speech synthesis tasks like text-to-speech (TTS) and speech-to-speech translation. In principle, there are infinitely many spoken renditions of a piece of text, and there is no universally agreed upon definition of what makes an utterance “correct”. Thus, researchers rely heavily on human evaluation, more specifically listening tests, during their day-to-day development cycles. The most popular type of listening test is MOS (Mean Opinion Score), during which several annotators listen to audio segments and rate them on a Likert scale between 1 to 5 (examples of foundational studies that use it include Tacotron [1], Parallel WaveNet [2], or FastSpeech 2 [3]). Listening tests can produce reliable results [4], since humans usually excel at detecting speech quality, and the scheme can be adapted to the need of every specific task. But they are also impractical and expensive: recruiting and polling annotators increases the cost of running experiments, slows down model research, and makes it impossible to compare results across time and institutions. The problem gets exacerbated in the multilingual setup: it may be challenging for researchers to find speakers of languages that are neither spoken by many, nor geographically close to them. Ultimately, this hinders their progress, skews the literature towards high resource languages, and prevents them from engaging in heavily multilingual research. This comes in contrast to text-based generation tasks, such as Machine Translation, for which the research community has long adopted automatic metrics (such as BLEU [5] or more recently COMET [6] and BLEURT [7]) as a complement to human assessment.

To address these issues, there has been a growing interest in developing automatic metrics for speech synthesis, spearheaded by systems such as AutoMOS [8], MOSNet [9], or LDNet [10]. The idea is to cast quality evaluation as a regression or classification problem: these systems predict a quality score from an utterance, using past listening tests as a source of training data. The task is difficult because the target domain is complex, even in a monolingual setup: synthesis artifacts come in many forms and can affect all levels of speech production, including pronunciation, prosody, voice, and audio quality. And the task is getting harder over time [11]: as systems progress, the focus has shifted from obvious artifacts (e.g., robotic voices) to more subtle errors, such as inappropriate prosody or mispronunciations. Yet, the same problem that motivates the task—plagues its solution—data is expensive to collect, especially outside high-resource languages, and so existing studies tend to use relatively limited training and testing sets. Early MOS Predictors have been shown to be brittle when used out of domain [12], and there are few studies outside English and Chinese.

In this paper, we study MOS Prediction at scale, and in a massively multilingual setup. We introduce SQuId (Speech Quality Identifier) a speech quality detector based on mSLAM (multilingual Speech and LAnguage Model), a recently published pre-trained model [13]. SQuId is trained on over a million ratings, an order of magnitude more than most recent studies [9]. More importantly it is, to the best of our knowledge, the first massively multilingual model for MOS prediction: we trained the model on 42 locales1 and tested it on 65. For comparison, VoiceMOS, the most comprehensive benchmark to date, covers two locales only [14]. We describe our dataset, our model, and show that SQuId outperforms a competitive baseline based on SSL and VoiceMOS by up to 50.0%. Most improvements come from the additional supervised multilingual data, complemented by minor optimizations that target the multilingual case. We then conduct several studies to highlight the factors that contribute to MOS prediction quality in this massively multilingual, in-the-wild, setup. Key insights from our work include:

(i) Training one model on a diverse dataset consisting of data from many locales consistently outperforms the monolingual approach as a result of cross-locale transfer, an effect well known in NLP [15–18], ASR [19–21] and TTS literature [22, 23]. The most spectacular manifestation of this phenomenon is the model’s strong performance on zero-shot locales, where there is no labelled MOS prediction data. Cross-locale transfer allows us to increase the model’s language coverage dramatically and has significant implications on evaluation of multilingual speech synthesis.

(ii) We conduct analyses to understand the nature and mechanism of cross-locale transfer for MOS prediction. We demonstrate that locale diversity has a large influence on model’s performance during fine-tuning, but transfer is driven less by language similarity and more by the presence of language-agnostic phenomena (possibly including diversity of audio quality, voices and TTS systems) in the dataset. We highlight the role of para-lingual transfer, by which

---

1 Compared to language, locales take regional variation into account. For instance English is covered by five locales in our dataset: US English, UK English, Indian English, Nigerian English, and Australian English. Each of these variants should be rendered differently by a TTS engine.
training data in one locale improves performance in another for reasons orthogonal to linguistics.

(iii) Through ablation studies, we highlight the importance of modeling choices, including multilingual pre-training and model capacity on SQuId’s performance.

Additional Related Work. Automatic MOS prediction has a long history in the TTS literature [24]. In addition to the systems cited above, recent work includes [25–29], none of which tackle multilinguality. Our work directly builds upon [12], which experiments with wide range of pre-trained models. Our methodology is close, but we scale up the data (from about 30K to 1.3M samples), number of locales (from 2 to 65), model size (from 300M to 600M parameters), and contribute a novel analysis of multilinguality. Cross-lingual transfer and massively multilingual NLP have a rich history in MT [15–17] and pre-trained models [18, 30]. Authors have studied cross-lingual transfer for at least a decade in speech recognition [31–35] and TTS [22, 36, 37].

2. MULTILINGUAL MOS PREDICTION IN THE WILD

We wish to predict MOS Naturalness ratings for both human and synthetic speech. Broadly speaking, MOS Naturalness describes how human-like an utterance sounds. Our main resource is an in-house corpus, aggregating approximately 1.9 million ratings in 66 locales across 2,092 research and commercial annotation projects completed between January 2021 and March 2022. Most of the audio is generated by TTS systems including both concatenative and neural systems, and the annotators are asked to select a rating based on how natural or unnatural the samples sound. The test sets used for the evaluations are primarily focused on TTS applications such as virtual assistant responses, driving directions, book passages, and news articles; general text from web-crawled corpora is also used.

Let us now present SQuId’s overall performance. To validate our approach, we first ensure that it performs well on VoiceMOS 22, currently the main benchmark for MOS prediction. We then scale up the test set and analyze its performance on 65 locales. Throughout the section, we will compare SQuId to Big-SSL-MOS, a competitive baseline in the spirit of SSL-MOS [12]. We fine-tune w2v-BERT on the main VoiceMOS database with a regression objective, using a 600M parameters version of the pre-trained model to ensure a fair comparison [12] uses up to 317M). The model comes in two variants: English-only, and multilingual. The architecture and dataset of our w2v-BERT implementation are described in detail in [13].

| Table 1: MOS Prediction dataset statistics. |
|-------------------------------------------|
| Num. training utt. / systems / locales | 969,589 / 1,476 / 42 |
| Num. dev utt. / systems / locales | 34,042 / 1,474 / 42 |
| Num. test utt. / systems / locales | 381,323 / 605 / 65 |
| Ave. utterance duration (seconds) / num. ratings | 4.5s / 1.4 |

4. PERFORMANCE

Let us now present SQuId’s overall performance. To validate our approach, we first ensure that it performs well on VoiceMOS 22, currently the main benchmark for MOS prediction. We then scale up the test set and analyze its performance on 65 locales. Throughout the section, we will compare SQuId to Big-SSL-MOS, a competitive baseline in the spirit of SSL-MOS [12]. We fine-tune w2v-BERT on the main VoiceMOS database with a regression objective, using a 600M parameters version of the pre-trained model to ensure a fair comparison [12] uses up to 317M). The model comes in two variants: English-only, and multilingual. The architecture and dataset of our w2v-BERT implementation are described in detail in [13].
Table 2: Results on VoiceMOS’22, Main Track. Segment-level correlation with human ratings (Kendall Tau). The VoiceMOS’22 results were obtained from [14], except SSL-MOS which we recalculated to confirm the alignment.

| Model          | Kendall Tau |
|----------------|-------------|
| VoiceMOS ’22  | 0.690       |
| SSL-MOS       | 0.690       |
| MOSA-Net      | 0.621       |
| LDNet         | 0.599       |
| Top           | 0.730       |
| Median        | 0.698       |
| Last          | 0.562       |
| Big-SSL-MOS   | 0.701       |
| English       | 0.693       |
| Multilingual  | 0.702       |
| SQuId         | 0.606       |
| SQuIdDS       | 0.700       |
| VoiceMOS      | 0.050       |
| SDS+VMOS      | 0.025       |

Correctness Check: VoiceMOS’22 - Main Track. VoiceMOS is the main public benchmark for MOS prediction. The main track involves 4,974, 1,066, and 1,066 utterances for training, development, and test respectively, all in English. The submissions are aggressively optimized towards the task; for instance UT-MOS [41], a top performing system, ensembles several models and use contextual information such as Listener and Domain ID. Additionally, the organizers provide the results of three recently published baselines: SSL-MOS [12], MOSA-Net [42], and LDNet [10]. Since the dataset is small comparatively (we test on two orders of magnitude more data in 30X more locales), we use it as a correctness check for our approach and baselines. We restrict ourselves to simple approaches: we use batch size 8 (instead of 32) and train for 10K steps (instead of a million), but do not introduce any additional optimization. Table 2 presents the results. We present three versions of SQuId: fine-tuned on the SQuId Dataset (SQuIdDS), VoiceMOS, and both sequentially. Although our approach is simple, we confirm that it is competitive. SQuId models trained on VoiceMOS data outperform all three VoiceMOS baselines, and would sit in the upper half of the benchmark. Our baseline Big-SSL-MOS performs even better, justifying that the comparison is fair.

Main Results. Figure 1 reports SQuId’s performance on the main dataset. We compute the correlation with human ratings on all 65 test locales and average the results. SQuId VoiceMOS is slightly better than Big-SSL-MOS English; the main difference is due to the switch from w2v-BERT to mSLAM, which performs better on this task. Using the SQuId Dataset incurs a dramatic 38% performance boost, which validates the usefulness of the dataset. The optimizations (re-sampling and locale identifier) add a more modest 3.75% increase. Our final model outperforms the best baseline by 50.0%, with superior results on 60 locales out of 65.

Figure 2 presents SQuId’s performance on each locale, as we increase the number of locales used for fine-tuning. Each addition improves the model, on both fine-tuned and zero-shot splits, demonstrating the existence of cross-locale transfer. SQuId’s performance is comparable in average on both sets, which illustrates the viability of zero-shot inference for MOS prediction.

There are large disparities between the scores, which can range from below 0 to over 0.3. A first, and possibly main, explanation, is that the test set is not homogeneous: the sub-tasks have various degrees of complexity and the annotations have various degrees of noisiness. Indirect evidence is that Big-SSL-MOS and SQuId tend to perform similarly on the same locales even though they were fine-tuned on very different datasets (Pearson correlation: 0.845). Furthermore, the training data is also heterogeneous, and so we cannot expect its performance to be constant, even on a perfect test set.

5. UNDERSTANDING CROSS-LOCALE TRANSFER DURING FINE-TUNING

Measuring Cross-Locale Transfer. To understand cross-locale transfer further, we focus on 8 languages and compare three models: mSLAM fine-tuned on the target locale only, on all 8 locales, and on all the locales we have (42). We picked languages from different families, with different levels of coverage (from 2.2K to 280K ratings), and for which we have ample test data to reduce the variance of the experiment. Figure 3 presents the results. For five out of eight setups, each addition increases the results. The most spectacular improvement happen between 1 and 8 on Thai (+244%) and Japanese (+31.2%). In contrast, adding data makes no difference or causes a slight degradation for three setups. We conclude that cross-locale “works” but it does not affect all locales uniformly.
Cross-locale transfer and language. Which locales are impacted by transfer? Intuitively, we may expect it to occur between tasks that are linguistically similar to each other [15, 20, 21]. To test this hypothesis, we focus on two languages for which we have different locales: Portuguese (the data set contains utterances in Brazilian and European Portuguese) and French (we have French and Canadian). We first fine-tune mSLAM on the locale with the lowest amount of data (i.e., European Portuguese and French, with about 9,700 examples each) and test it on the same. We then compare mSLAM fine-tuned on locales pairs, by coupling the target locale with either a close one (e.g., European Portuguese and Brazilian Portuguese) or remote ones (e.g., European Portuguese and Japanese). If the hypothesis holds, then adding in-language data will help, and similar pairs will perform better than remote ones. We chose the baselines locales such that the amount of fine-tuning data is approximately similar in all setups. Figure 4 presents the results. There is much transfer from Brazilian to European Portuguese, but the effect is approximately similar with German. Furthermore, Japanese and Arabic are not far behind. The results are more spectacular on French: adding French Canadian to European French actually decreases its performance, while Icelandic improves it. Our conclusions are therefore negative: it seems that the influence of linguistic similarity on cross-lingual transfer is very limited in this application. A similar observation may be found in [22], where improvements are observed in TTS systems transferring voices across highly dissimilar languages like English and Chinese.

Cross-Lingual vs. Para-Lingual Transfer. If language similarity does not facilitate transfer, then what does? We hypothesize that the task is in fact dominated by sound and paralinguistics, i.e., artifacts such as robotic voices and flat prosody that do not depend on words. This does not mean that there is no cross-locale transfer: the performance on each task does indeed improve as the model is exposed to the others. But the phenomenon is distinct from the cross-lingual transfer that is commonly discussed in the NLP literature [15]. We distinguish para-lingual transfer from cross-lingual transfer, and treat them as two orthogonal components of cross-locale transfer.

To highlight the role of para-lingual transfer, we fine-tune single-locale mSLAM models on the 8 locales discussed above and run them on each other’s test set, i.e., we run each of the 8 models on each of the 8 locales. We present the results in a matrix where the rows represent fine-tuning and columns test locales (Figure 5). We trained randomly initialized models to prevent any cross-lingual transfer, with a smaller architecture to stabilize the learning (we used the 42M parameters model discussed in the next section). If the tasks were truly independent, the diagonal would be salient. If they depended on language alone, there would be more affinity between languages of the same family than between very different ones, or at least, between languages that share phonological features. Instead we find that the results are diffuse: the good results of Turkish on Portuguese could come from a shared phonemic inventory, but the fact that Japanese yields relatively good results on five out of eight tasks (doing better than French on French) confirms that MOS Naturalness is, to some extent, orthogonal to language.

Transfer between Pre-Training and Fine-Tuning If the task is dominated by para-linguistics and sound artifacts, does it still make sense to pre-train models in multiple languages, and do we still benefit from scale? We rerun our experiments varying the sets of pre-training languages (from 1 to 51) and capacities (from 42M to 600M parameters), using speech-only models trained on VoxPopuli [43] to simplify the setup.

6. CONCLUSION

Zero-shot MOS prediction seems to be a viable path towards lowering the cost of massively multilingual speech synthesis evaluation. However, cross-locale transfer may in fact have little to do with language similarity. Our study holds lessons for both TTS evaluation and SSL. For TTS evaluation, we see that naturalness is only one facet of speech quality, and reiterate the importance of using complementary evaluation methods, both human and automatic, to measure language-specific aspects such as prosody and pronunciation. Regarding SSL, it seems plausible that para-locale transfer affects other tasks well, such as TTS and ASR. We therefore call for model developers to enrich their pre-training corpora to include a diverse set of voice (natural and synthetic), prosody, and sound [44], in addition to variations in domains and languages.
7. REFERENCES

[1] Skerry-Ryan RJ Wang, Yuxuan et al., “Tacotron: Towards end-to-end speech synthesis.” Proc. Interspeech, 2017.

[2] Aaron van den Oord, Yazhe Li, et al., “Parallel WaveNet: Fast high-fidelity speech synthesis,” in Proc. ICML, 2018, pp. 3918–3926.

[3] Yi Ren, Chenxu Hu, et al., “Fastspeech 2: Fast and high-quality end-to-end text to speech.” Proc. ICLR, 2021.

[4] James R Lewis, “Psychometric properties of the mean opinion scale,” Proceedings of HCI International 2001: Usability Evaluation and Interface Design, pp. 149–153, 2001.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. ACL, 2002, p. 311–318.

[6] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie, “COMET: A neural framework for MT evaluation,” in Proc. EMNLP, 2020.

[7] Thibault Sellam, Dipanjan Das, and Ankur Parikh, “BLEURT: Learning robust metrics for text generation,” in Proc. ACL, 2020.

[8] Brian Patton, Yannis Agiomyrgiannakis, et al., “AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech;” in NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop, 2016.

[9] Chen-Chou Lo, Szu-Wei Fu, et al., “MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion,” in Proc. Interspeech 2019, pp. 1541–1545.

[10] Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda, “LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech,” in ICASSP 2022, 2022, pp. 896–900.

[11] Sébastien Le Maguer, Simon King, and Naomi Harte, “Back to the Future: Extending the Blizzard Challenge 2013,” in Proc. Interspeech, 2022, pp. 2378–2382.

[12] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi, “Generalization ability of MOS prediction networks,” in ICASSP 2022, 2022, pp. 8442–8446.

[13] Ankur Banpa, Colin Cherry, et al., “mSLAM: Massively multilingual joint pre-training for speech and text,” arXiv, 2022.

[14] Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi, “The VoiceMOS Challenge 2022.” Interspeech, 2022.

[15] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight, “Transfer learning for low-resource neural machine translation,” arXiv, 2016.

[16] Roeo Aharoni, Melvin Johnson, and Orhan Firat, “Massively multi-lingual neural machine translation,” in Proc. NAACL, 2019, pp. 3874–3884.

[17] N. Arivazhagan, Ankur Banpa, Orhan Firat, et al., “Massively multilingual neural machine translation in the wild: Findings and challenges,” ArXiv, 2019.

[18] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al., “Unsupervised cross-lingual representation learning at scale,” in Proc. ACL, 2020, pp. 8440–8451.

[19] Hui Lin, Li Deng, Dong Yu, Yi-fan Gong, Alex Acero, and Chin-Hui Lee, “A study on multilingual acoustic modeling for large vocabulary ASR,” in Proc. ICASSP, 2009, pp. 4333–4336.

[20] Bo Li, Tara N Sainath, et al., “Multi-dialect speech recognition with a single sequence-to-sequence model,” in Proc. ICASSP, 2018.

[21] Anjuli Kannan, Arindrima Datta, et al., “Large-scale multilingual speech recognition with a streaming end-to-end model,” arXiv, 2019.

[22] Yu Zhang, Ron J. Weiss, et al., “Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning.” ArXiv, 2019.

[23] Mutian He, Jingzhou Yang, Lei He, and Frank K Soong, “Multilingual Byte2Speech models for scalable low-resource speech synthesis,” arXiv, 2021.

[24] Alessandro Ragano, Emmanuel Benetos, et al., “A comparison of deep learning mos predictors for speech synthesis quality,” arXiv, 2022.

[25] Jennifer Williams, Joanna Rownicka, Pilar Oplustil, and Simon King, “Comparison of speech representations for automatic quality estimation in multi-speaker text-to-speech synthesis,” ArXiv, 2020.

[26] Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin, “MBNET: MOS prediction for synthesized speech with mean-bias network,” in ICASSP, 2021, pp. 391–395.

[27] Wei-Cheng Tseng, Chien-yu Huang, Wei-Tsung Kao, Y st Y Lin, and Hung-yi Lee, “Utilizing self-supervised representations for mos prediction,” Interspeech ‘21, 2021.

[28] Yeunjoo Choi, Youngmoon Jung, and Hoirin Kim, “Neural MOS prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification,” in IEEE SLT Workshop, 2021, pp. 462–469.

[29] Xiaohai Tian, Kaiqi Fu, Shaojun Gao, Yiwei Gu, Kai Wang, Wei Li, and Zejun Ma, “A Transfer and Multi-Task Learning based Approach for MOS Prediction,” in Proc. Interspeech 2022, 2022, pp. 5438–5442.

[30] Linting Xue, Noah Constant, et al., “mT5: A massively multilingual pre-trained text-to-text transformer,” in Proc. NAACL-HLT, 2021.

[31] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, “Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR,” in IEEE SLT, 2012, pp. 246–251.

[32] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in Proc. ICASSP, 2013, pp. 7304–7308.

[33] Xinjian Li, Siddharth Dalimia, et al., “Universal phone recognition with a multilingual allophone system,” in Proc. ICASSP, 2020.

[34] Piotr Zelasko, Laureano Moro-Velázquez, et al., “That sounds familiar: an analysis of phonetic representations transfer across languages,” arXiv, 2020.

[35] Siyuan Feng, Piotr Zelasko, et al., “How phonotactics affect multilingual and zero-shot ASR performance,” in Proc. ICASSP, 2021.

[36] H. Javkin, K. Hata, et al., “A multilingual text-to-speech system,” in Proc. ICASSP, 1989, pp. 242–245.

[37] Ariadna Sanchez, Alessio Falai, et al., “Unify and conquer: How phonetic feature representation affects polyglot text-to-speech (TTS) in Interspeech, 2022.

[38] Rasmus Dall, Junichi Yamagishi, and Simon King, “Rating naturalness in speech synthesis: The effect of style and expectation,” in Proceedings of Speech Prosody. Citeseer, 2014.

[39] Mirjam Wester, Cassandra Valentini-Botinhao, and Gustav Eje Henter, “Are we using enough listeners? no! — an empirically-supported critique of Interspeech 2014 TTS evaluations,” in Proc. Interspeech, 2015, pp. 3476–3480.

[40] Qingsong Ma, Johnny Wei, Onurfe Bojar, and Yvette Graham, “Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges,” in Proc. WMT, Florence, Italy, 2019, pp. 62–90.

[41] Takaaki Saeki, Detai Xin, et al., “UTMOS: Utokyo-sarulab system for MOS Prediction,” in Proc. Interspeech, 2022, pp. 5438–5442.

[42] Ryandhimas E Zezario, Szu-Wei Fu, et al., “Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features,” in Proc. ICASSP, 2021, pp. 391–395.

[43] Changhan Wang, Morgane Riviere, et al., “VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation,” in Proc. ACL-IJCNLP, 2021.

[44] Wei-Ning Hsu, Anuroop Sriram, et al., “Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training,” arXiv, 2021.