I. Results across languages

In this supplementary material, we provide additional results from the MOS in Table 8 in the main paper. We present the naturalness MOS values across all the target speaker - input language combinations, as defined in Table 6 in the main paper. Here, **en-ch** corresponds to English target speaker, and input language as Chattisgarhi - this corresponds to cross-lingual synthesis. On the other hand, **en-en** shows the results with English target speaker, and input language as English. The **bold blue** values and the blue values are the best and second scores across each speaker-language combination, respectively. The teams are ranked as per Table 8 in the main paper. The discussion on these results are present below.

A. Track 1

Table 1 shows the target speaker and language specific results for Track 1. The scores show that the higher-ranked teams have achieved consistently high scores across all synthesis combinations. TEAM 1 demonstrates superior performance, achieving the best scores in 6 combinations and second-best scores in the remaining 3 combinations. Lower-ranked teams show inconsistent performance across languages, with TEAM 8 exhibiting notable variability - achieving high scores in select pairs (en-ch, kn-bn, kn-mr) but very low scores in others (en-en, en-kn, hi-en). The lower scores could indicate insufficient or improper training of the TTS model, and the variability could indicate problems with how speaker and text information is combined.

Analysis of the languages reveals an interesting pattern in same-language versus cross-language synthesis. Among same-language combinations (en-en, hi-hi, kn-kn), only hi-hi shows consistently high performance across top teams. However, this pattern does not extend to other same-language pairs in top teams - cross-lingual synthesis often obtains higher scores compared to same-language synthesis. For instance, TEAM 1's performance in cross-language pairs (en-ch: 4.67, kn-mr: 4.58) exceeds their same-language scores (en-en: 4.38, kn-kn: 4.12). This indicates that the TTS models have effectively learned language-agnostic speaker characteristics.

TABLE 1. Naturalness MOS across target speaker and language combination for track 1

Team	en-ch	en-en	en-kn	hi-en	hi-hi	hi-te	kn-bn	kn-kn	kn-mr
TEAM 1	**4.67**	4.38	**4.45**	**4.08**	4.55	**4.50**	**4.67**	4.12	**4.58**
TEAM 2	3.42	3.92	3.25	**4.08**	4.53	3.67	4.17	3.71	4.08
TEAM 3	3.67	3.71	2.92	**4.08**	**4.61**	3.58	4.25	4.00	4.00
TEAM 4	3.50	3.46	3.75	3.25	4.18	3.33	3.08	**4.29**	2.91
TEAM 5	3.25	**4.58**	2.25	**4.08**	3.50	3.00	2.42	2.92	3.42
TEAM 6	3.58	3.67	2.00	3.50	2.96	2.42	3.00	3.21	2.75
TEAM 7	3.25	3.58	2.50	2.92	3.41	2.25	2.42	2.83	3.42
TEAM 8	4.08	1.75	1.42	1.00	3.24	3.58	4.00	2.81	4.25
TEAM 9	1.92	2.62	2.91	2.92	2.96	1.92	2.17	3.12	2.50

TABLE 2. Naturalness MOS across language and target speaker combination for track 2

Team	en-ch	en-en	en-kn	hi-en	hi-hi	hi-te	kn-bn	kn-kn	kn-mr
TEAM 10	4.50	**4.75**	**4.08**	**4.92**	4.30	**4.17**	4.33	**4.21**	3.67
TEAM 2	**4.58**	3.96	3.75	3.92	**4.43**	3.64	4.08	3.96	3.55
TEAM 1	4.17	3.46	3.92	3.50	4.33	3.67	**4.42**	4.12	**4.25**
TEAM 5	3.92	4.62	2.42	4.17	4.00	2.50	2.25	3.10	3.33
TEAM 7	3.17	3.58	2.64	2.92	3.35	2.64	1.92	3.48	2.92
TEAM 6	3.67	3.42	2.50	3.33	2.95	1.92	3.17	2.29	3.42
TEAM 9	2.58	2.50	2.27	2.67	2.52	1.67	2.75	3.04	2.33

B. Track 2

Table 2 shows the target speaker and language specific results for Track 2. Track 2 allowed the use of external datasets for training TTS model, and provided target speaker files for few-shot training. Team 10 has performed significantly well on the naturalness MOS on most of the pairs. While top teams maintained consistently high performance similar to Track 1, Track 2's allowance of external TTS datasets revealed interesting insights about data availability's impact on performance. This is particularly evident in TEAM 2's results across both tracks, where they showed improvement only in English language pairs (en-ch, en-en, en-kn) but experienced performance degradation in other languages. This pattern strongly suggests that while the availability of large external datasets like VCTK and LibriTTS benefited English synthesis, the lack of comparable large multi-speaker corpora for Indian languages limited potential improvements in those language pairs.

For English target speaker (en-), we see that en-kn combinations, compared to en-ch and en-en, consistently showed lower scores across teams, suggesting particular challenges with this language pair. For Hindi target speaker (hi-), there's notable success in both same-language (hi-hi) and English input (hi-en) combinations, with TEAM 10 achieving an impressive 4.92 for hi-en, the highest score across all combinations. However, hi-te combinations showed consistently lower scores, indicating specific challenges in Hindi-Telugu synthesis. For Kannada target speaker (kn-), there's considerable variation across input languages - while kn-bn and kn-mr showed relatively good scores among top teams, the same-language synthesis (kn-kn) didn't necessarily outperform cross-lingual combinations. This language-specific analysis suggests that success in cross-lingual synthesis depends not just on the availability of training data, but also on potentially inherent linguistic similarities or differences between language pairs and the target speaker.

C. Track 3

Table 3 shows the target speaker and language specific results for Track 3. Track 3 allowed the use of external training data, but the synthesis was zero-shot for the target speaker. Similar to other tracks, the top teams have consistently performed well on all most of the combinations, with TEAM 2 consis-

tently achieving high naturalness scores. While TEAM 10 and TEAM 1 also showed strong results, their performance displayed more variation across different language pairs. TEAM 2 has better performance when compared to it's track 2 scores, but the reasoning cannot be attributed to zero-shot nature of the track as the team used a different TTS model in track 3.

TABLE 3. Naturalness MOS across language and target speaker combination for track 3

Team	en-ch	en-en	en-kn	hi-en	hi-hi	hi-te	kn-bn	kn-kn	kn-mr
TEAM 2	4.17	4.46	**4.09**	**4.58**	**4.68**	**4.45**	4.00	**4.55**	4.00
TEAM 10	4.42	**4.58**	3.67	4.33	4.60	3.83	3.75	4.00	3.92
TEAM 1	**4.58**	3.96	3.67	3.83	3.64	2.83	**4.33**	3.96	**4.75**
TEAM 11	3.17	3.96	2.10	3.42	3.00	2.25	3.27	3.04	2.92
TEAM 6	3.75	3.46	2.58	3.00	2.95	2.75	3.42	2.78	3.42
TEAM 5	2.50	4.29	2.00	3.83	3.24	1.92	2.00	2.25	2.09
TEAM 9	2.25	1.87	2.25	2.00	2.76	2.25	2.73	2.62	2.08

II. Overall findings across tracks

We find that same-language synthesis wasn't necessarily better than cross-lingual synthesis. This indicates the speaker representations in these models are capable of being language-agnostic. The use of external datasets primarily benefited English language pairs, and limited improvement in Indian languages due to lack of large multi-speaker corpora. There are challenges related to specific language pair, indicating the need for more understanding on linguistic similarity between languages for building multilingual models. Overall the top teams have consistently performed well, and lower teams have either performed poorly or shown variation in their scores across different speaker-language pairs.