# Houston we have a Divergence:
# A Subgroup Performance Analysis of ASR Models

Alkis Koudounas, Flavio Giobergia

Politecnico di Torino

{firstname.lastname}@polito.it

## I. ABSTRACT

The Fearless Steps APOLLO Community Resource [1] provides an unparalleled opportunity to explore the vast potential of historical multi-speaker team communications from the NASA Apollo missions. This study focuses on discovering which are the characteristics that make Apollo recordings more or less intelligible to Automatic Speech Recognition (ASR) methods.

We extract, for each audio recording, interpretable metadata on the recordings themselves (signal-to-noise ratio, spectral flatness, presence of pauses, duration of the sentence), on the transcript (number of words spoken, speaking rate) or known *a priori* (speaker). Next, we identify subgroups of audio recordings based on combinations of the extracted metadata (e.g., {*speaker=NEIL, snr=low, total_duration=low*} to indicate short, noisy recordings of Neil Armstrong). We compute the performance (e.g., Word Error Rate) for each subgroup and the difference in performance (*"divergence"*) w.r.t the overall population. We rely on the work presented in [2] for the extraction of relevant subgroups and their divergences.

We then apply the Whisper model [3] in different sizes (`base`, `small`, `medium`, `large-v3`), trained on English-only or multilingual datasets[1], in zero-shot or after fine-tuning. We conduct several analyses to (i) automatically identify and describe the most problematic subgroups for a given model, (ii) examine the impact of fine-tuning w.r.t. zero-shot at the subgroup level, (iii) understand the effect of model size on subgroup performance, and (iv) analyze if multilingual models are more sensitive than monolingual to subgroup performance disparities.

*a) Subgroup Performance Assessment:* We report subgroup-specific performance metrics in Table I to identify subgroups with the poorest ASR performance for the `medium-en` and `medium-multilingual` models. This analysis informs us about the challenges posed by different subpopulations within the Fearless Steps APOLLO corpus. We additionally report in Figure 1(a) the Global Shapely Values [2], [4] for the `base-en` model. This is informative of the type of metadata that affects positively or negatively the predictions. For example, recordings with a high SNR and a low spectral flatness are generally associated with better performance (i.e., lower WER), as intuition dictates. Other information about the speakers indicates the extent to which various people are intelligible, according to the ASR model.

*b) Pre-trained vs. fine-tuned models:* We investigate the impact of fine-tuning on ASR models by comparing the performance of pre-trained models with their fine-tuned counterparts. Table II shows that fine-tuning consistently globally improves ASR performance. In addition, we show that all divergences get closer to 0: for negative divergences, this means that the fine-tuned model improves on the subgroups that performed poorly in zero-shot). For positive divergences, it means that the fine-tuned model does not improve as much for the already well-performing subgroups. In other words, fine-tuning reduces the divergence of the model across subgroups.

*c) Base vs. large models:* We compare the performance of the base and the small ASR models in a zero-shot setting. It has been shown in literature that larger models do not necessarily consistently outperform smaller models across all subgroups [5], [6]. In the first block of Table III, we analyze the different performance between the `base-en` and the `small-en` models. The smaller model outperforms the larger one in some subgroups, even if with a relatively modest gap. Conversely, the improvement observed when transitioning from the base to the small model is substantially higher.

*d) Multilingual vs. English-only Models:* We examine the performance disparity between multilingual and English-only ASR models, identifying subgroups for which the multilingual model consistently outperforms the English-only counterpart. For example, Figure 1(b) shows the variation in performance between the subgroups, when passing from English-only to multilingual. Interestingly, most subgroups benefit from using the English-only model. However, a small number of subgroups actually decrease in performance. Examples of such subgroups are {*n_words=high, snr=high, speaker=FD1, trim_dur=high*} and {*speaker=FD1, spectral_flatness=low, tot_dur=high*}

In conclusion, our study provides a detailed comparative analysis of ASR methods applied to the Fearless Steps APOLLO corpus. The insights gained enhance our understanding of subgroup-specific performance variations, paving the way for advancements in the development and optimization of ASR systems for Earth-to-space communications.

---

[1]The large model has only been pre-trained on the multilingual dataset

| Model | Subgroups | $WER$ | $\Delta_{WER}$ | $t$ |
|---|---|---|---|---|
| medium-en | $\mathtt{S}^-$: {*snr=high, spectral_flatness=medium, speakRate=high, speakRate_trim=high, trim_dur=medium*} | 82.985 | 2.985 | 2.41 |
| | $\mathtt{S}^+$: {*speaker=CAPCOM1, spectral_flatness=low, speakRate_trim=high*} | 76.703 | -3.297 | 2.71 |
| medium-multi | $\mathtt{S}^-$: {*n_words=medium, snr=medium, speaker=FD1*} | 85.219 | 2.853 | 2.92 |
| | $\mathtt{S}^+$: {*speaker=BUZZ, spectral_flatness=medium, speakRate_trim=high*} | 79.830 | -2.536 | 2.63 |

TABLE I: WER divergence for the most negatively ($\mathtt{S}^-$) and positively ($\mathtt{S}^+$) divergent subgroups compared to overall test performance. The $t$ column indicates the Welch's t-test value.
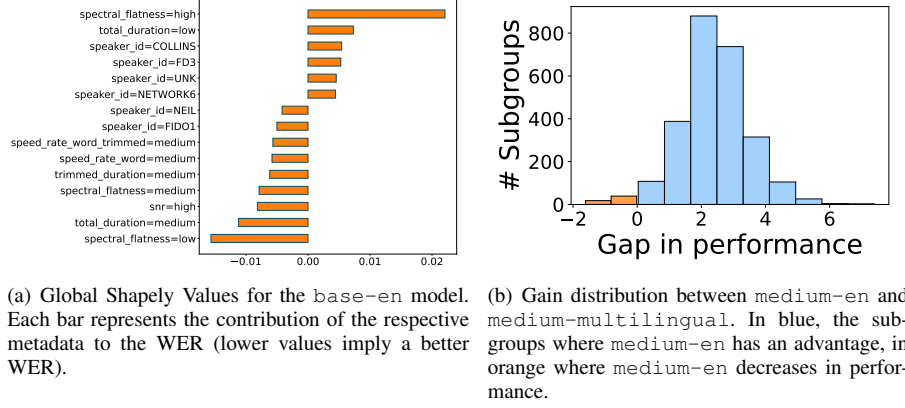


(a) Global Shapely Values for the base-en model. Each bar represents the contribution of the respective metadata to the WER (lower values imply a better WER).

(b) Gain distribution between medium-en and medium-multilingual. In blue, the subgroups where medium-en has an advantage, in orange where medium-en decreases in performance.

Fig. 1: (**Left**) Global Shapely Values for a single model, (**Right**) Gain distribution between two models.

| Model | $WER$ | $\Delta_{WER}^{min}$ | $\Delta_{WER}^{max}$ | $\Delta_{WER}^{avg}$ | $\Delta_{WER}^{std}$ |
|---|---|---|---|---|---|
| base-en | 89.997 | -1.528 | 1.299 | 0.015 | 0.411 |
| base-ft | 77.444 | -1.077 | 0.983 | -0.002 | 0.297 |
| small-en | 85.009 | -2.051 | 2.134 | -0.057 | 0.532 |
| small-ft | 69.975 | -0.934 | 0.790 | 0.004 | 0.229 |
| medium-en | 80.000 | -3.297 | 2.985 | -0.050 | 0.797 |
| medium-ft | 60.028 | -0.637 | 0.883 | -0.011 | 0.197 |
| large-v3 | 75.024 | -1.964 | 1.921 | 0.026 | 0.528 |
| large-v3-ft | 49.996 | -1.657 | 1.667 | -0.029 | 0.385 |

TABLE II: Information about the Word Error Rate for various models, either fine-tuned or not. Performance reported overall, and in terms of subgroup divergence. For the latter, the spread, minimum, maximum, and average divergences are reported.

| Comparison | | Subgroups | $gap_{WER}$ | $f_{M1}$ | $f_{M2}$ | $t$ |
|---|---|---|---|---|---|---|
| base-en vs. small-en | ↓ | {*snr=high, trim_dur=low*} | 1.04 | 88.76 | 87.80 | 1.99 |
| | ↑ | {*speaker=MSTC1, trim_dur=low*} | -4.67 | 90.71 | 86.05 | 2.21 |
| medium-en vs. medium-multi | ↓ | {*snr=medium, speakRate_trim=medium, trim_dur=low*} | 7.4 | 76.94 | 84.34 | 2.64 |
| | ↑ | {*n_words=high, snr=high, spkID=FD1, trim_dur=high*} | -1.61 | 82.12 | 80.51 | 2.06 |
| large-v3 vs. large-v3-ft | ↑ | {*n_words=low, spkID=FD1, speakRate=medium, speakRate_trim=medium*} | -28.09 | 76.94 | 48.86 | 2.30 |

TABLE III: WER Performance gap when changing the model size (base vs. small), the pre-training objective (medium-en vs. medium-multi), or the training methodology (zero-shot or fine-tuned). (↑) denotes the highest performance improvement (a lower WER indicates a better score), (↓) indicates the largest decrease.

REFERENCES

[1] J. H. Hansen, A. Joglekar, S.-J. Chen, M. C. Shekar, and C. Belitz, "Fearless steps apollo: Advanced naturalistic corpora development," in *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022*, 2022.

[2] E. Pastor, L. De Alfaro, and E. Baralis, "Looking for trouble: Analyzing classifier behavior via pattern divergence," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 1400–1412.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[4] A. Koudounas, F. Giobergia, and E. Baralis, "Bad exoplanet! explaining degraded performance when reconstructing exoplanets atmospheric parameters," in *NeurIPS 2023 AI for Science Workshop*, 2023. [Online]. Available: https://openreview.net/forum?id=9Z4XZOhwiz

[5] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, L. Cagliero, L. de Alfaro, E. Baralis, and D. Amberti, "Exploring subgroup performance in end-to-end speech models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[6] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, E. Reale, L. Cagliero, S. Cumani, L. de Alfaro, E. Baralis, and D. Amberti, "Towards comprehensive subgroup performance analysis in speech models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.