OPEN ACCESS

Short Report

Hearing aids in the era of foundation models

Hörgeräte im Zeitalter der Grundmodelle

Abstract

The recent introduction of foundation models (FMs) has taken the world by storm. Ranging from large language models (LLMs) to image and audio analysis and generation, FMs have introduced a new paradigm in artificial intelligence (AI), one where practitioners transition from standard supervised machine learning to prompting and in-context learning. This has implications for hearing aid research, and specifically for the use of such models for noise attenuation and speech enhancement. Even though the uptake of FMs is minimal to non-existent for this application domain, mainly due to the prohibitive computational complexity of those models, there are nevertheless ways to benefit from FM advances in an indirect way. We review these approaches in the present contribution.

Zusammenfassung

Die jüngste Einführung von Grundmodellen (FMs) hat die Welt im Sturm erobert. Von großen Sprachmodellen (LLMs) bis hin zur Analyse und Generierung von Bild- und Audiodateien haben FMs einen Paradigmenwechsel in der künstlichen Intelligenz (KI) hervorgerufen, bei dem Anwender vom herkömmlichen überwachten maschinellen Lernen zu Textanfragen und kontextbezogenem Lernen übergehen. Dies hat ebenfalls Auswirkungen auf die Hörgeräteforschung, insbesondere auf die Verwendung solcher Modelle zur Geräuschunterdrückung und zur Verbesserung der Sprachqualität. Obwohl die Anwendung von FMs in diesen Kontext bisher minimal bis nicht existent ist, hauptsächlich aufgrund der prohibitiven Rechenkomplexität der Modelle, gibt es dennoch Möglichkeiten, von den Fortschritten durch FMs auf indirekte Weise zu profitieren. Wir überprüfen diese Ansätze in dem vorliegenden Beitrag.

Andreas Triantafyllopoulos[1,2]
Björn W. Schuller[1,2,3,4]

1 CHI – Chair of Health Informatics, Technical University of Munich, MRI, Munich, Germany

2 MCML – Munich Center for Machine Learning, Munich, Germany

3 MDSI – Munich Data Science Institute, Munich, Germany

4 GLAM – Group on Language, Audio & Music, Imperial College, London, United Kingdom

Introduction

Hearing aids aim to compensate for hearing loss by processing the input audio stream and manipulating it in such away so as to partially recover lost hearing. While recovering hearing covers multiple facets of the human experience, such as being able to partake in conversations or enjoy music, recovering the ability to understand human speech is understandably one of the main priorities for hearing aid devices. Their key operating principles leverage advances in a wide array of fields, from physics, to electronics, (psycho)acoustics, digital signal processing (DSP), statistics, and – increasingly – artificial intelligence (AI) [1]. In particular, AI features prominently as a complement, or even substitute, to DSP components [2], primarily the ones tackling noise reduction and attenuation [3], [4]. In a new frontier, *foundation models* (FMs) have appeared as a novel class of models in the broader AI community [5], but have not yet found their way into hearing aid research. Foundation models (FMs) differ from traditional deep neural networks (DNNs) in that they exhibit *emergent properties*, i.e., capabilities that they were not explicitly trained to perform but that can be uncovered through the successful use of *prompting* [6].

Prompts can be thought of as a mixture of *cues* and *instructions* provided to a model. Instructions pertain to the task that should be solved; cues add additional context that can be leveraged to improve performance. For instance, a large language model (LLM) might be asked to classify the sentiment of a target sentence ("The weather is nice today."). On top of the sentence to be classified, the input query must be prefaced with an instruction ("Predict the sentiment of the following sentence.") and can be further constrained according to the specifications of the user ("Predict the sentiment of the following sentence. Select one from positive, negative, neutral. Answer in one word.").

Auditory FMs operate on similar principles as LLMs, albeit with audio as a primary or secondary input [7]. Text prompts now become audio prompts. The input may be

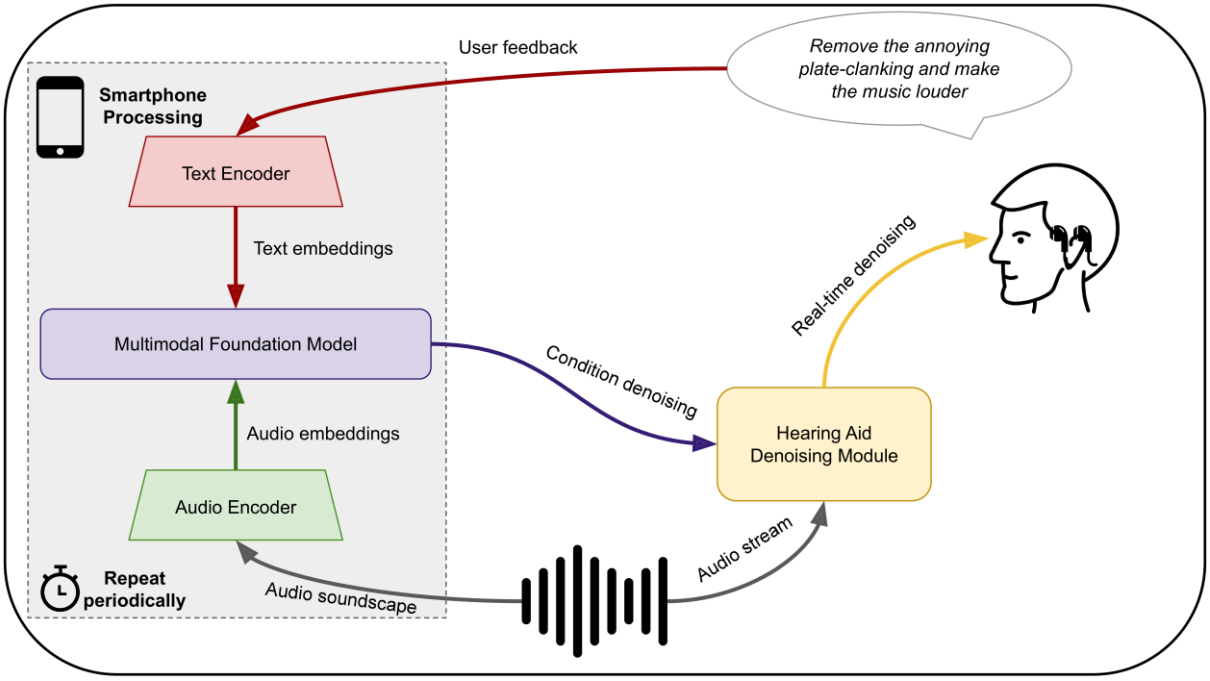

Figure 1: Schematic overview of how multimodal foundation models can be used in conjunction with hearing aids. The auditory component of the FM continuously monitors the background audio, while the textual component receives user feedback. The FM processes both types of inputs periodically and sends information to the hearing aid, which uses it to adapt its denoising process.

an audio snippet while instructions are typically left as text; this provides an intuitive interface for downstream users. A typical application is audio manipulation (e.g., for denoising, inpainting, or voice transformation). The input query corresponds to an *[AUDIO]* snippet, with *[AUDIO]* symbolising a discretised and compressed representation of the audio that needs to be manipulated by the foundation model. This is then prefaced/followed by a textual instruction and additional cues. For example, the input may be: "The following audio consists of one target male speaker and background music. Remove all background music and keep all target speech intact. *[AUDIO]*." Importantly, this interleaving between audio and language (or even other modalities) relies in mapping all inputs to a joint embedding space that the FM can process. As most existing off-the-shelf FMs are LLMs, this is typically achieved by adding an audio module that process the input *[AUDIO]*, followed by a mapping module that translates the output of the audio module to the space shared by text (i.e., *tokens)*, and the entire prompt is fed into an FM for further processing. Similarly, the output is treated as an audio stream (and may or may not be fed into additional audio modules for decoding).

Foundation models for hearing aids

Having reviewed the basic principles of FMs, we now turn to the critical question of how they can be used in the context of hearing aids. At first glance, computational complexity is an obvious factor prohibiting their uptake. FMs employ billions of model parameters [5], with the "small" versions of those models typically featuring 7-billion parameters. Even at the most extreme level of quantisation currently possible for AI models (4-bits), this still results in 3.5 GB of memory just to load the model, without accounting for the storage of intermediate computations, or, indeed, the runtime to pass an input through the model. Obviously, the deployment of such a model in a hearing aid is a long time away. However, there are ways to leverage FMs which circumvent the hardware showstopper discussed above. The key insight lies in offloading those compute-intensive models to external devices. Given the proliferation of smartphones and their integration in the hearing-aid ecosystem, as well as the emergence of new, distributed sensing paradigms like "Auracast" [8], there are nowadays complementary devices that can record and process audio with significantly higher compute capabilities than hearing-aid devices [9].

An overview of the process is shown in Figure 1. In a nutshell, FMs are employed to do what they excel at – general world understanding, which is then co-opted to improve denoising performance. The key motivation for using FMs like that is that the world is slowly-changing, at least in the terms relevant for a hearing aid user. Coupled with the fact that most people nowadays carry a smartphone connected to the Internet, this allows for offloading the running of the FM to a device outside the hearing aid. This can relay the necessary information – essentially a model of the surrounding environment – from the FM back to the hearing aid, which can then utilise that information to improve its signal processing. While exotic at first glance, this procedure can enable us to leverage the advances in FMs without waiting for accompanying improvements in hardware. In the next two

sections, we conceptualise how that might become possible, beginning with an introduction of audio FMs, and proceeding with a perspective on how they can be employed in hearing aid practice.

Naturally, running these models externally introduces an additional latency that precludes online usage. However, there lies immense potential in their ability to understand the underlying environment even in an offline setting. In particular, the ultimate goal of a hearing aid is hearing loss compensation, which, when it comes to speech understanding, is partially achieved through speech enhancement and noise attenuation. The latter is contingent on the type of noise prevalent. Oftentimes, this noise is quasi-stationary, as in the typical examples of babble noise, restaurants, transportations, or music. These types of noises change slowly – slowly enough that a large foundation model only needs to sense them sporadically (e.g., every few seconds or even minutes). They can be applied on periodic recordings of the environment to generate a detailed characterisation of it which can be provided to the hearing aid and condition its denoising algorithm.

Examples of this type of conditioning have already proven successful for general speech denoising [10], whereby a *fingerprint* of the background noise is used as additional information to improve noise attenuation. This fingerprint is processed by a separate encoder – which, in principle, can be more complex than the main branch as it only needs to be run rarely – and its output is used for the conditioning of a main denoising network. While previous works have used standard neural networks for this fingerprint encoder, performance can be largely improved by relying on the more advanced class of FMs now available. Similarly, this process can be used to enrol the target speaker to be enhanced – a form of personalisation that is well-known in the literature.

Beyond automatically understanding the background audio type, however, FMs can be used to foster a more intuitive and adaptive interaction with the user of the hearing aid. As mentioned, auditory FMs can seamlessly combine audio with linguistic queries – the latter can be provided in real time by the user, who could dynamically adjust the parameters of the hearing aid to match their current need. We note that such “profiles” are already available as part of smartphone apps that allow for the configuration of a hearing aid – however, the use of descriptive, natural language can provide a more timely and granular adaptation, as well as introduce a trial-and-error component, with the user iterating through queries.

In summary, we expect FMs to gradually make their way into the next generations of hearing aids as supplements that run on external devices. They have the capacity to serve as a powerful sidekick to the speech enhancement and denoising capabilities of hearing aids, thus paving the way for better hearing loss compensation.

Notes

Conference presentation

This contribution was presented at the 26th Annual Conference of the German Society of Audiology.

Competing interests

The authors declare that they have no competing interests.

References

1. Dillon H. Hearing aids. New York: Thieme Medical Publishers Inc.; 2001.
2. Wang D. Deep Learning Reinvents the Hearing Aid: Finally, wearers of hearing aids can pick out a voice in a crowded room. IEEE Spectr. 2017 Mar;54(3):32-7. DOI: 10.1109/MSPEC.2017.7864754
3. Hamacher V, Chalupper J, Eggers J, Fischer E, Kornagel U, Puder H, Rass U. Signal processing in high-end hearing aids: State of the art, challenges, and future trends. EURASIP Journal on Advances in Signal Processing. 2005;18:1-15. DOI: 10.1155/ASP.2005.2915
4. Schröter H, Rosenkranz T, Escalante-B AN, Maier A. Low latency speech enhancement for hearing aids using deep filtering. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2022;30:2716-28. DOI: 10.1109/TASLP.2022.3198548
5. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Niladri C, Chen A, Creel K, Davis JQ, Dorottya D, Demszky, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Koh PW, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Yuhuai W, Xie SM, Yasunaga M, You J, Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K, Liang P. On the Opportunities and Risks of Foundation Models. ArXiv. 2021:arXiv:2108.07258. DOI: 10.48550/arXiv.2108.07258
6. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi EH, Hashimoto T, Vinyals O, Liang P, Dean J, Fedus W. Emergent Abilities of Large Language Models. Transactions on Machine Learning Research. 2022 Jun 26;(08):1-30. DOI: 10.48550/arXiv.2206.07682
7. Liu H, Chen Z, Yuan Y, Mei X, Liu X, Mandic D, Wang W, Plumbley MD. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. MLResearchPress; 2023. p. 21450-74

8. Bluetooth Market Development. An Overview of Auracast™ Broadcast Audio. Bluetooth SIG, Inc.; 2024 [last accessed 2024 Sep 16]. Available from: https://www.bluetooth.com/bluetooth-resources/overview-of-auracast-broadcast-audio/

9. Kaufmann TB, Foroogozar M, Liss J, Berisha V. Requirements For Mass Adoption Of Assistive Listening Technology By The General Public. In: IEEE, editor. Proceedings of the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). Piscataway, NJ: IEEE; 2023. p. 1-5. DOI: 10.1109/ICASSPW59220.2023.10193566

10. Liu S, Keren G, Parada-Cabaleiro E, Schuller B. N-HANS: A neural network-based toolkit for in-the-wild audio enhancement. Multimed Tools App. 2021 Jul;80(6):28365-89. DOI: 10.1007/s11042-021-11080-y

Corresponding author:
Andreas Triantafyllopoulos
Department for Clinical Medicine, TUM School of Medicine and Health, Klinikum rechts der Isar (Public Sector Institution), Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany
andreas.triantafyllopoulos@tum.de

Please cite as
Triantafyllopoulos A, Schuller BW. Hearing aids in the era of foundation models. GMS Z Audiol (Audiol Acoust). 2024;6:Doc28. DOI: 10.3205/zaud000063, URN: urn:nbn:de:0183-zaud0000639

This article is freely available from
https://doi.org/10.3205/zaud000063

Published: *2024-12-17*