Bi-modal architectures for deeper user preference understanding from spoken content

Moreno La Quatra
moreno.laquatra@polito.it
Politecnico di Torino
Turin, Italy

Lorenzo Vaiani
lorenzo.vaiani@polito.it
Politecnico di Torino
Turin, Italy

Luca Cagliero
luca.cagliero@polito.it
Politecnico di Torino
Turin, Italy

Paolo Garza
paolo.garza@polito.it
Politecnico di Torino
Turin, Italy

Abstract

The analysis of spoken content often relies on a two-stage architecture, which comprises (i) audio-to-text conversion and (ii) transcript processing. However, discarding additional information hidden in audio signals could result in losing the underlying information that can be used for improving text-based analysis and user engagement. Being more correlated with emotional features, audio features could unveil hidden user preferences towards specific multimedia content. This paper elaborates on practical use cases in which audio-related content can be processed jointly with textual data to better capture user preferences.

Keywords: Multimodal Learning, Speech Processing, Spoken Language Understanding

1 Introduction

In the latest years, Natural Language Understanding has seen a steadily increasing interest from different research communities. Deep learning-based language models (e.g., [5], [2]) have shown to be very effective in modelling the semantics behind natural language. State-of-the-art solutions leverage the generalization capabilities of deep neural networks to learn actionable audio signal representations [9]. Despite the majority of the past solutions focused on the analysis of audio transcriptions, the recent diffusion of internet-based platforms offering direct access to spoken content (e.g., podcasts or audio books) has fostered the analysis of raw audio signals. Specifically, several creators focused on the production of spoken audio content for their user-bases. This resulted in an increasing availability of large audio collections, which can be easily accessed by several millions of users [3]. Unlike musical content, spoken language recordings are usually composed of two complementary information streams, i.e., natural language and audio. We argue that natural language content and audio play a different roles in the deep understanding of user engagement and preferences. While the speech transcription could reveal semantically relevant information, audio processing is instrumental for detecting, for instance, the mood of the speaker, the level of engagement, and the underlying user preferences.

A robust analysis of spoken content requires the simultaneous processing of both audio and textual modalities. This paper envisions a bi-modal architecture for spoken content representation that can leverage both NLP and audio processing.

1.1 Opportunities

Spoken content understanding offers interesting opportunities. Unlike other multimedia content (e.g., videos), audio tracks cannot leverage visual attention to attract users. Despite this could seem a weakness, it may turn into a possible strength: users consume pure audio content in multiple scenarios while their visual attention is engaged in other activities (e.g. driving or manual working).

The widespread diffusion of audio tracks has resulted in the evolution of several research lines related to spoken content understanding. Several approaches rely on the extraction of textual transcripts from the audio samples, which are then processed using NLP techniques. However, audio speech tanscription is not error-free and misses important information: the speaker tone of voice, feeling, and empathy. Moreover, environmental features captured by the audio recordings are also missing. We claim that the study of the user engagement and sentiment can be relevantly improved by learning deep models from multimodal sources combining both transcript and raw audio.

1.2 Challenges

The convergence between NLP and Computer Vision architectures towards the design of *attention-based* models [5, 6] prompts their joint application in the audio domain, where newly proposed architectures pose major focus on the speech recognition task [9]. The intrinsic nature of audio data requires specific efforts in the design of attention-based models to effectively cope with raw audio signals.

Similar to hand-written textual documents, audio signals may convey speech in multiple languages. To disentangle

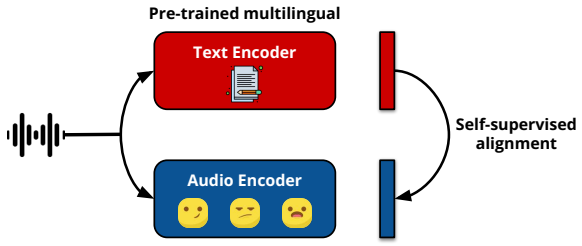

Figure 1. Sketch of the multimodal framework.

emotional features from the discussed topic, we seek new, language-agnostic architectures. The training of models capable of mapping audio segments in a multilingual latent space is therefore one of the main challenges in this field.

2 Multimodal framework

Lately, the audio processing community has mainly focused on the design of Deep Learning architectures for Automatic Speech Recognition (ASR) [1]. However, as discussed above, audio representations derived from the transcripted text are incomplete and potentially biased.

An effective architecture for multimodal analysis should be able to disentangle the expressive content of audio from the semantic content of text. It can leverage two different encoder networks that are able to cope with specific modalities: audio and text encoders. Figure 1 illustrate a sketch of the high-level architecture.

2.1 Training data and procedures

While several benchmark collections for spoken content are already available [4, 7], they often contain content in different languages. The proposed high-level architecture could leverage pre-trained multilingual text models [8] to bootstrap the audio encoder branch. By leveraging self-supervised techniques (e.g., contrastive learning), the text encoder can serve as a guide for the audio branch. Self-supervised alignment can be used to generate language-independent audio representations. Specific fine-tuning phases using parallel annotated data could be effective for disentangling the two modalities.

2.2 User-centric applications

Unified data representations could be effective in multiple application scenarios. Considering the increasing interest of large user-bases towards podcast content, it is possible to leverage bi-modal representations in the design of advanced recommendation models. Text encoders should be able to model user preferences towards specific topics whereas the underlying audio features can be exploited to reveal the most relevant user characteristics (e.g., the speaker tone and timbre of voice).

Similar considerations hold in the domain of spoken content summarization (e.g., political debates). While delivering similar semantic content, different parts of an audio recording could convey different emotional charges and be more engaging than others, thus being more suitable to be part of a summary. Finally, it is possible leverage on audio-related features to better detect user satisfaction for voice-based customer assistance. While the conversation transcription could highlight the topic and possible negative/positive verbal feedback, the speaker's voice tone could incorporate valuable information about the actual user engagement.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020), 12449–12460.

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[3] Ben Carterette, Rosie Jones, Gareth F. Jones, Maria Eskevich, Sravana Reddy, Ann Clifton, Yongze Yu, Jussi Karlgren, and Ian Soboroff. 2021. *Podcast Metadata and Content: Episode Relevance and Attractiveness in Ad Hoc Search.* Association for Computing Machinery, New York, NY, USA, 2247–2251. https://doi.org/10.1145/3404835.3463101

[4] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics.* International Committee on Computational Linguistics, Barcelona, Spain (Online), 5903–5917. https://www.aclweb.org/anthology/2020.coling-main.519

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations.* https://openreview.net/forum?id=YicbFdNTTy

[7] Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 8229–8233. https://doi.org/10.1109/ICASSP40776.2020.9054626

[8] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 4512–4525.

[9] Luyu Wang, Pauline Luc, Yan Wu, Adria Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, Joao Carreira, et al. 2021. Towards Learning Universal Audio Representations. *arXiv preprint arXiv:2111.12124* (2021).