# Review 157: [Short] End-to-End Speech Recognition Contextualization with Large Language Models, 30.09.23

**Paper: https://arxiv.org/abs/2309.10917v1**

https://huggingface.co/papers/2309.10917

אודיו וטקסט? נכון ששני סוגי דאטה אלו הם די שונים ולא הגיוני להניח שמודל שאומן על טקסט יכול להביא תוצאות טובות גם על אודיו לאחר כיול קל. אך התברר שזה אפשרי. במאמר שנסקור היום ב-#shorthebrewpapereviews למעשה לקחו מודל שפה מאומן והשתמשו בו בשביל לבצע משימה audio2text. כלומר להפיק את מה שנאמר בקטע אודיו.

איך הם עשו זאת? מכיוון שאי אפשר סתם לקחת אות אודיו להזין אותו כמו שהוא למודל שפה נדרש כאן אנקודר שמקודד את הפיצ'רים המהותיים של אות אודיו. מחברי המאמר משתמשים במודל מאומן מראש הנקרא ConFormer ומפיק לנו ייצוג לטנטי של אות אודיו (כלומר מערך של וקטורים המייצגים כל מקטע של אודיו או בפשטות טוקני אודיו). ד״א ConFormer הוא מודל די מעניין (הוצע על ידי גוגל) המשלב ארכיטקטורת הטרנספורמר עם שכבות קונבולוציה (משתמשים שם גם בקידוד מיקום יחסי RoPE שנהיה מאוד פופולרי היום).

לאחר מכן לוקחים את ייצוג של טוקני האודיו ומזינים אותם למודל שפה מאומן (הם לקחו LLAMA) יחד עם עוד מידע על האודיו כמו שם הוידאו שממנו הוא נלקח או התיאור הטקסטואלי. בסוף מטייבים(fine-tune) מודל שפה בסגנון LoRA על דאטהסט המורכב מזוגות של אודיו והטקסט. ולהפתעתי זה עובד ממש לא רע.