

זיהוי דיבור אוטומטי: החזית הבאה

התפתחויות, הזדמנויות והשפעות של טכנולוגיית זיהוי דיבור אוטומטי בתחומים שונים

טל רוזנוין

פוסט זה מתמקד בהתקדמות בטכנולוגיית זיהוי דיבור אוטומטי (ASR) והשפעתה על תחומים שונים. ASR הפך נפוץ בתעשיות מרובות, עם שיפור הדיוק כתוצאה של הגדלת גודל המודל וגודל מערכי הנתונים (מתוייגים ולא מתוייגים) לאימון המודלים.

במבט קדימה, טכנולוגיית ASR צפויה להמשיך ולהשתפר עם הגדלת המודל האקוסטי ושיפור מודל השפה הפנימי. בנוסף, טכניקות אימון כגון פיקוח עצמי (self supervised) וריבוי משימות (multi-task) יפרצו דרך לשימוש ב-ASR גם בשפות בהן יש קושי למצוא דאטה זמין בכמויות. יתר על כן, אימון רב-לשוני יאפשר קפיצת מדרגה נוספת בביצועים, ויאפשר שימוש מוצרי בסיסי כגון פקודות קוליות בשפות אלה.

ASR גם ימלא תפקיד משמעותי ב-Generative AI, שכן האינטראקציה עם אוטרים תהיה באמצעות ממשק אודיו/טקסט. עם הופעתו של NLP ללא טקסט (textless NLP), כמה משימות קצה, כגון תרגום מבוסס דיבור לדיבור (speech to speech translation), עשויות להיפתר ללא שימוש במודל ASR מפורש. מודלים מולטי-מודאליים המקבלים כקלט טקסט, אודיו או שניהם גם יחד ישוחררו, ויאפשרו לחולל טקסט או לסנתז אודיו כפלט.

יתר על כן, מערכות דיאלוג בעלות ממשק אדם-מכונה מבוסס קול ישפרו את עמידותן בפני שגיאות תמלול והבדלים בין צורות כתובות ומדוברות. זה יספק עמידות למבטאים מאתגרים ולדיבור של ילדים, מה שיאפשר לטכנולוגיית ASR להפוך לכלי חיוני עבור יישומים רבים.

להערכתנו, תשחרר מערכת אחודה לשיפור דיבור-ASR-דיאריזציה מקצה לקצה, שתאפשר התאמה אישית של מודלי ASR, ותאפשר תמלול איכותי בדיבור חופף ובתרחישים אקוסטיים מאתגרים. זהו צעד משמעותי לקראת פתרון האתגרים של טכנולוגיית ASR בתרחישים בעולם האמיתי.

לבסוף, צפוי גל של ממשקי API של ASR. אף על פי כן ישנן הזדמנויות לסטארט-אפים קטנים להתעלות על חברות טכנולוגיה גדולות בתחומים עם יותר מגבולות חוקיות או רגולטוריות על השימוש בטכנולוגיה/רכישת נתונים, ובאוכלוסיות עם שיעורי אימוץ טכנולוגיה נמוכים.

שנת 2022, סקירה

טכנולוגיית זיהוי דיבור אוטומטי (ASR) תופסת תאוצה בתעשיות שונות כגון חינוך, פודקאסטים, מדיה חברתית, רפואה טלפונית, מוקדים טלפוניים ועוד. דוגמה מצוינת היא השכיחות הגוברת של ממשק אדם-מכונה (HMI) מבוסס-קול במוצרי צריכה, כגון מכונות חכמות, בתים חכמים, טכנולוגיה מסייעת חכמה [1], סמארטפונים ואפילו עוזרי בינה מלאכותית (AI) בבתי מלון [2]. על מנת לענות על הדרישה ההולכת וגוברת לתגובתיות מהירה ומדויקת, מודלי ASR עם זמן תגובה מהיר פותחו למשימות כמו איתור מילות מפתח [3], איתור סיום דיבור [4] ותמלול [5]. דגמי ASR המיוחסים לדובר [6-7] זוכים גם הם

לתשומת לב מכיוון שהם מאפשרים התאמה אישית של המוצר, ומספקים ערך רב יותר למשתמשי הקצה.

שכיחות נתונים. פלטפורמות המאפשרות הזרמת אודיו ווידאו כגון מדיה חברתית ויוטיוב הובילו לרכישה קלה של נתוני אודיו ללא תיוג [8]. טכניקות חדשות בפיקוח עצמי הוצגו כדי להשתמש באודיו הזה מבלי להזדקק לתיוג [9-10]. טכניקות אלו משפרות את הביצועים של מערכות ASR בדומיין היעד, אפילו ללא כוונן עדין של המודל על נתונים מסומנים באותו תחום [11]. גישה נוספת שזוכה לתשומת לב בשל יכולתה לנצל את הנתונים הלא מסומנים הללו היא אימון עצמי באמצעות פסאודו תיוג [12-13]. הרעיון הוא לתמלל אוטומטית מקטעי אודיו ללא תוויות באמצעות מערכת זיהוי דיבור אוטומטי (ASR) ולאחר מכן להשתמש בתמלול שנוצר כאילו נוצר על ידי אדם, לצורך אימון מערכת ASR אחרת בצורה ממוקדת. OpenAI נקטה בגישה שונה, בהנחה שהם יכולים למצוא באינטרנט תמלולים שנוצרו על ידי אדם בקנה מידה. הם יצרו דאטהסט איכותי ובקנה מידה גדול (K640 שעות) לאימון על ידי הורדת מקטעי אודיו זמינים לציבור עם כתוביות שנוצרו על ידי גורם אנושי. באמצעות מערך נתונים זה, הם אימנו מודל ASR (הידוע גם בשם Whisper) באופן ממוקד לחלוטין, תוך השגת תוצאות בחזית הטכנולוגיה (SoTA) במספר מבחנים מקובלים, ללא כיוון נוסף של המודל [14].

פונקציות מטרה. למרות שאימון מקצה לקצה (End-2-End- E2E) זו הפרדיגמה השולטת באימון מודלי ASR [15-17], פונקציות מטרה חדשות עדיין מתפרסמות. הוצגה טכניקה חדשה הנקראת מתמר אוטומטי-רגרסיבי היברידי (HAT) [18], המאפשרת למדוד את האיכות של מודל השפה הפנימית (ILM) על ידי הפרדת ההסתברות הא-פוסטריורית של התו הריק (blank) משאר התווים. עבודה מאוחרת יותר [19] השתמשה בפקטוריזציה זו כדי לעדכן ביעילות את ה-ILM תוך שימוש בנתונים טקסטואליים בלבד, מה ששיפר את הביצועים הכוללים של מערכות ASR, במיוחד את התמלול של יישויות, מונחי סלנג ושמות עצם, שהם נקודות כאב עיקריות עבור מערכות ASR. בנוסף, פותחו מדדים חדשים על מנת להתיישר טוב יותר עם התפיסה האנושית ולהתגבר על בעיות סמנטיות של מדדים קיימים - שיעור שגיאות מילים (WER) [20].

בחירת ארכיטקטורת מודל. לגבי הבחירות של מבנה המודל האקוסטי, Conformer [21] נשאר מועדף עבור דגמי סטרימינג, בעוד טרנספורמרים [22] היא ארכיטקטורת ברירת המחדל עבור דגמים שאינם סטרימינג. לגבי האחרון, מודלים רב-לשוניים של מקודד בלבד (wav2vec [23-24]) ומקודד-מפענח (Whisper [14]) הוצגו. מודלים אלה שיפרו ביצועים על פני מספר מבחנים ביחס לתוצאות ה-SoTA. מודלים אלה מתגברים על מקביליהם הסטרימינגים בשל גודל המודל, גודל בסיס הנתונים לאימון, והקונטקסט הגדול אותו הם רואים.

פיתוחי בינה מלאכותית רב-לשונית של ענקיות הטכנולוגיה. גוגל הכריזה על "יוזמת 1,000 שפות" לבניית מודל בינה מלאכותית התומך ב-1,000 השפות המדוברות ביותר [25], בעוד ש-Meta AI הכריזה על מאמציה ארוכי הטווח לבנות כלים לתרגום אוטומטי (MT) הכוללים את רוב שפות העולם [26].

פריצת דרך בשפה מדוברת. שוחררו מודלי מקודד-מפענח רב-מודאליים (דיבור/טקסט) שיועדים לבצע מספר משימות כגון 5SpeechT [27], המראים הצלחה רבה במגוון רחב

של משימות עיבוד שפות מדוברות, כולל ASR, סינתזת דיבור, תרגום דיבור, המרת קול, שיפור דיבור וזיהוי דובר.

התקדמות אלו בטכנולוגיית ASR צפויות להניע חדשנות נוספת ולהשפיע על מגוון רחב של תעשיות בשנים הבאות.

מבט לעתיד

למרות האתגרים, תחום זיהוי הדיבור האוטומטי (ASR) צפוי לעשות התקדמות משמעותית בתחומים שונים, החל ממודלים אקוסטיים וסמנטיים ועד בינה מלאכותית שיחתית וגנרטיבית (Conversational AI and Generative AI), ואפילו ASR מוטה דובר. חלק זה מספק הצצה מפורטת לתחומים אלה ומשתף את התחזיות שלי לעתיד טכנולוגיית ASR.

שיפורים כלליים:

השיפור של מערכות ASR צפוי הן בחלקים האקוסטיים והן בחלקים הסמנטיים.

בנוגע למודל האקוסטי, הגדל המודל וגודל בסיס הנתונים לאימון צפויים לשפר את הביצועים הכוללים של מערכות ASR, בדומה להתקדמות הנצפית במודל שפה גדולים (LLMs). למרות ששינוי קנה המידה של מקודדים בלבד (כגון Wav2Vec או Conformer) מהווה אתגר, צפויה פריצת דרך שתאפשר זאת, או לחילופין נראה מעבר לארכיטקטורות מקודד-מפענח דוגמת Whisper. עם זאת, לארכיטקטורות מסוג זה יש חסרונות שיש לטפל בהם, כגון הזיות. אופטימיזציות כגון Faster-Whisper [28] ו-2NVIDIA-wav2vec [29] יפחיתו את זמן האימון ופרדיקציה בפרודקשן, ויפחיתו את המחסום למבצוע של מודלי ASR גדולים.

בצד הסמנטי, החוקרים יתמקדו בשיפור מודלים של ASR על ידי הגדלת ההקשר האקוסטי או הטקסטואלי. הזרקת טקסט לא מצומד (לאודיו) בקנה מידה גדול ל-ILM במהלך אימון E2E, כמו ב-JEIT [30], תיבדק גם כן. מאמצים אלה יסייעו להתגבר על אתגרים מרכזיים כגון תמלול מדויק של ישויות עם שם, מונחי סלנג ושמות עצם.

למרות שמודל הדיבור האוניברסלי (USM) של גוגל [31] ו-whisper שיפרו את ביצועי מערכת ה-ASR בהשוואה למספר מבחנים, עדיין יש צורך לפתור כמה מבחנים שכן שיעור השגיאה (WER) נותר סביב 20% [32]. שימוש במודלי דיבור יסודיים (foundation models), הוספת דאטה מגוון יותר ויישום למידה מרובה משימות ישפרו משמעותית את הביצועים במבחנים אלה, ויפתחו הזדמנויות עסקיות חדשות. יתרה מכך, מדדים חדשים צפויים להתפרסם כדי לחדד מטרות ולשפר ביצועים במשימות קצה שעד כה פחות נחקרו, כגון צילי שיחה לא לקסיקליים [33] בתחום הרפואי וזיהוי וסיווג מילות מילוי [34] בעריכת מדיה ותחומים חינוכיים. מודלים מכוונים ספציפיים למשימה עשויים להתפתח למטרה זו. לבסוף, עם הצמיחה של ריבוי-מודאליות, צפויים להשתחרר מודלים נוספים, מערכי נתונים ומבחנים חדשים עבור מספר משימות קצה [35-36].

ככל שההתקדמות נמשכת, צפוי גל של ממשקי API ל ASR, בדומה לעיבוד שפה טבעית (NLP) USM. של גוגל, Whisper של OpenAI ו-1-Conformer של Assembly [37] הם חלק מהסוגיות הראשונות.

למרות שזה נשמע טיפשי, עימוד בכוח (force alignment) עדיין מאתגר עבור חברות רבות. קוד קוד פתוח עבור זה עשוי לעזור לרבים להשיג יישור מדויק בין קטעי אודיו לתמלול המתאים להם.

שפות עם מעט דאטה זמין:

התקדמות בלמידה בפיקוח עצמי, למידה מרובה משימות ומודלים רב-לשוניים צפויים לשפר את הביצועים בשפות דלות משאבים ושפות בלתי כתובות באופן משמעותי. שיטות אלו ישיגו ביצועים משביעי רצון על ידי שימוש במודלים שאומנו מראש בצורה לא מפוקחת, וכיוונו בצורה עדינה על מספר קטן יחסית של דגימות מתווגות [24]. גישה מבטיחה נוספת היא dual learning, למידה כפולה [38], פרדיגמה ללמידת מכונה מפוקחת למחצה המבקשת למנף נתונים לא מתווגים על ידי פתרון שתי משימות הפוכות (טקסט לדיבור (TTS) ו-ASR במקרה שלנו) בבת אחת. בשיטה זו, כל מודל מייצר פסאודו תווית עבור דוגמאות ללא תווית, המשמשות לאימון המודל השני.

בנוסף, שיפור ILM באמצעות טקסט לא מצומד יכול לשפר את ביצועי המודל במיוחד במשימות קצה עם לקסיקון סופי וידוע מראש כגון פקודות קוליות. עבור יישומים מסויימים, כגון כתוביות לסרטוני YouTube, הביצועים יהיו משביעי רצון אך לא מושלמים. ביישומים אחרים, כגון יצירת תמלול מילולי בבית המשפט, זה עשוי לקחת עוד זמן עד שהביצועים יעברו בתנאי הסף. אנו צופים שחברות יאספו נתונים על סמך מודלים אלה תוך תיקון ידני של תמלילים בשנת 2023, ונראה שיפורים משמעותיים בשפות עם משאבים נמוכים לאחר כוונן עדין על נתונים קנייניים בשנת 2024.

AI גנרטיבי:

השימוש באוטוארים צפוי לחולל מהפכה באינטראקציה האנושית עם נכסים דיגיטליים. בטווח הקצר, ASR ימשש כאחד היסודות ב- Generative AI שכן האוטורים הללו יתקשרו באמצעות ממשק טקסטואלי/שמיעתי.

אבל בעתיד, שינויים עשויים להתרחש כאשר תשומת הלב עוברת לכיווני מחקר חדשים. לדוגמה, טכנולוגיה מתפתחת שסביר שתאומץ היא Textless NLP, המייצגת גישה מידול חדשה למידול שפה לטובת גנרציה של אודיו [39]. גישה זו משתמשת ביחידות אודיו בדידות הניתנות ללמידה [40], ויוצרת באופן אוטומטי את יחידת האודיו הבדידה הבאה יחידה אחת בכל פעם, בדומה לדרך בה בחוללי טקסט עובדים. לאחר שאותן יחידות דיסקרטיות חוללו, הן ניתנות לפריסה בחזרה לתחום האודיו. עד כה, טכנולוגיה זו הצליחה לייצר המשך דיבור עם ביצועים סבירים מבחינה תחבירית וסמנטית, תוך שמירה על זהות הדובר והפרוזודיה עבור דוברים שלא נראו בזמן האימון, כפי שניתן לראות ב-GSLM/AudioLM [39, 41]. הפוטנציאל של טכנולוגיה זו הוא עצום, שכן ניתן לדלג על רכיב ה-ASR (והשגיאות שלו) במשימות רבות. לדוגמה, שיטות תרגום מסורתיות של תרגום מבוסס דיבור לדיבור (speech to speech translation) פועלות באופן הבא: הן מתמללות את האמירה בשפת המקור, לאחר מכן מתרגמות את הטקסט לשפת היעד באמצעות מודל תרגום מכונה, ולבסוף מייצרות את האודיו בשפות היעד באמצעות מנוע TTS. באמצעות טכנולוגיית textless NLP, ניתן לבצע s2s translation באמצעות ארכיטקטורת מקודד-מפענח יחיד הפועלת ישירות על יחידות אודיו בדידות מבלי להשתמש במודל ASR מפורש [42]. אנו צופים כי מודלים עתידיים של Textless NLP יפתרו משימות רבות כגון מערכות תשובות לשאלות, מבלי לעבור מפורשות דרך תמלול. עם זאת, החיסרון העיקרי של שיטה זו

הוא מעקב אחר שגיאות וניטור בעיות, מכיוון שהדברים יהיו פחות אינטואיטיביים כאשר עובדים על מרחב היחידות הבדידות במקום עבודה על התמלול.

מודלי 5T [43] ו-OT [44] הראו הצלחה רבה ב-NLP תודות לאימון מרובה המשימות שלהם. בשנת 2021 פורסם 5SpeechT [27] שהראה הצלחה רבה במשימות שונות של עיבוד שפה מדוברת. מוקדם יותר השנה, VALL-E [45] ו-VALL-EX [46] שוחררו. הם הראו יכולות למידה מרשימות בתוך הקשר עבור מודלי TTS על ידי שימוש בטכנולוגיית textless NLP, המאפשרת חיקוי של קול הדובר על ידי שימוש במקטע אודיו של הדובר באורך מספר שניות, וללא צורך בכיוון עדין. בנוסף, ניתן אפילו לדובר את הדובר במספר שפות.

על ידי שילוב השיטות של 5SpeechT ו-VALL-E, אנו יכולים לצפות לשחרור של מודלים דמויי OT שיודעים לקבל כקלט טקסט, אודיו או שניהם, ולחולל טקסט או לסנתז אודיו כפלט, בהתאם למשימה. עידן חדש של מודלים יתחיל, שכן למידה בתוך הקשר תאפשר הכללה למשימות חדשות ללא אימון. זה יאפשר חיפוש סמנטי על אודיו, תמלול דובר יעד באמצעות מקטע אודיו קצר של אותו הדובר או תיאורו בטקסט חופשי, למשל, "מה הילד הצעיר שהשתעל אמר?". יתר על כן, זה יאפשר לנו לסווג או לסנתז אודיו באמצעות אודיו או תיאור טקסטואלי ולפתור משימות NLP ישירות מאודיו באמצעות ASR מפורש/מרומז.

Conversational AI:

בינה מלאכותית לשיחה (conversational AI) מופיעה בעיקר בדמות מערכות דיאלוג מוכוונות משימות (personal assistant-PA), כגון Alexa של אמזון ו-Siri של אפל. מערכות אלה הפכו פופולריות בשל יכולתן לספק גישה מהירה לפונקציונאליות ולמידע באמצעות פקודות קוליות. בעוד שענקיות הטכנולוגיה שולטות בשוק, תקנות רגולציה חדשות יאלצו אותן להציע אפשרויות של צד שלישי לשירותים מסוג זה, ויפתחו את השוק לתחרות [47]. כאשר זה יקרה, אנו צופים התממשקות ותקשורות בין PAs. יכולת זו תאפשר להשתמש בכל מכשיר כדי להתחבר לכל PA בכל מקום בעולם [48]. מנקודת המבט של ASR, זה יציב אתגרים חדשים שכן ההקשר יהיה הרבה יותר רחב, וה PAs יצטרכו להיות עמידים למבטאים שונים ואף לתמוך במספר שפות ודיאלקטים.

במהלך השנים האחרונות, חלה קפיצת מדרגה טכנולוגית בביצועי מערכות דיאלוג פתוח מבוססות טקסט, למשל, בלנדר-בוט ו-LaMDA [49–50]. במקור מערכות אלה עבדו בדומיין הטקסטואלי בלבד - הן מקבלות טקסט כקלט ומחוללות טקסט במוצא. ככל שביצועי ASR השתפרו, מערכות דיאלוג פתוח הונגשו עם HMI מבוסס קול, מה שהוליד אתגרים חדשים הנובעים מהבדלים בין הצורה המדוברת והכתובה כגון קושי בזהוי ישויות, ושגיאות תמלול כגון שגיאות הגייה [51–52].

פתרונות אפשריים יגיעו משיפור איכות התמלול, וממודלים שפה חזקים שיתמודדו ביהיו חסינים לשגיאות תמלול והגייה. מדד ביטחון (confidence score) עבור תחזית המודל האקוסטי [53] יכול להיות שחקן מפתח, כיוון שיאפשר לזהות שגיאות דובר או לשמש כקלט נוסף למודל השפה או לוגיקת הדקודינג. יתרה מזאת, אנו מצפים שמודלי ASR ינבאו רמזים לא מילוליים כגון סרקזם, ויאפשרו לסוכנים להבין את השיחה בצורה מעמיקה יותר ולספק תגובות טובות יותר.

שיפורים אלו יאפשרו לדחוף מערכות דיאלוג נוספות עם HMI אודיטורי כדי לתמוך במבטאים מאתגרים ובדיבור של ילדים, כמו ב-Loora [54] ו-Speaks [55].

בהסתכלות שאפתנית עוד יותר, אנו מצפים לשחרור של מסגרת למידה מקצה לקצה מרובת משימות, עבור משימות שפה מדוברת תוך שימוש במודלים משותפים של בעיות דיבור ו-NLP [56]. מודלים אלו יתאמנו בצורה E2E שתפחית את השגיאה המצטברת בין מודלים עוקבים וייתחסו למשימות כגון הבנת השפה המדוברת, סיכום מדובר, ותשובות לשאלות מדוברות, על ידי הזנת דיבור כקלט ופליטת פלטים שונים במוצא כגון תמלול, ישויות עם שם, סיכומים ותשובות לשאלות טקסטואליות.

התאמה אישית (פרסונליזציה) מהווה גורם משמעותי עבור PAs ומערכות דיאלוג פתוח, מה שיובייל אותנו לנקודה הבאה: ASR מותנה דובר.

ASR מותנה דובר:

תמלול שיחות בהן הדוברים והדוברות רחוקים מהמקורפונים המקליטים היא משימה לא פתורה, אפילו מערכות מתקדמות (SoTA) יכולות להשיג רק כ-35% WER [57].

סנציות ראשונות של מודלים אחודים לתמלול ודיאריזציה פורסמו כבר בשנת 2019 [58]. השנה, אנו יכולים לצפות לשחרור של מערכת לשיפור דיבור-ASR-דיאריזציה מקצה לקצה אשר תשפר את הביצועים על מקטעי אודיו בהם יש דיבור חופף, ותאפשר עמידות רבה יותר לסיטואציות מאתגרות כגון הדהוד, תמלול מרחוק ותמלול ביחס אות-רעש (SNR) נמוך. השיפור יושג באמצעות אופטימיזציה משותפת של משימות, שיטות משופרות של אימון מקדים (כגון WavLM [10]), יישום שינויים בארכיטקטורת המודל [59], הגדלת בסיס הנתונים במהלך אימון מקדים, גם אם הדאטה איננו מתויג אך מדומיין המטרה [11]. יתרה מכך, אנו יכולים לצפות לשחרור של מערכות ASR המותנות דובר לטובת פרסונליזציה. זה ישפר עוד יותר את דיוק התמלול של קולו של דובר היעד ותטה את הטקסט המתומלל למילים המוגדרות על ידי המשתמש, כגון שמות אנשי קשר, שמות עצם, שהן חיוניות עבור PAs [60]. בנוסף, מודלים עם זמן תגובה נמוך ימשיכו להיות תחום מיקוד משמעותי כדי לשפר את החוויה הכוללת וזמן התגובה של מכשירי קצה [61-62].

תפקידים של סטארט-אפים למול ענקיות הטכנולוגיה:

למרות שענקיות הטכנולוגיה צפויות להמשיך לשלוט בשוק עם ה-API שלהן, סטארטאפים קטנים עדיין יכולים להתעלות בביצועיהם בדומיניים ספציפיים. לדוגמא, בדומיניים עם ייצוג נמוך בדאטהסטים לאימון של הענקיות עקב תקנות/רגולציות, דוגמת התחום הרפואי ודיבור של ילדים, ואוכלוסיות שעדיין לא אימצו טכנולוגיה, כמו מהגרים עם מבטאים מאתגרים או אנשים שלומדים ונשים שלומדות אנגלית ברחבי העולם. בנוסף, בשווקים שאינם גדולים דיים עבור ענקיות הטכנולוגיה, כמו שפות המדוברות במדינות קטנות, חברות סטארט-אפ קטנות עשויות למצוא הזדמנויות להצליח ולהניב רווח.

כדי ליצור מצב של win-win, ענקיות הטכנולוגיה יכולות לספק ממשקי API המציעים גישה מלאה לפלט של המודלים האקוסטיים שלהן, תוך שהן מאפשרות לאחרים לכתוב את לוגית הדקודינג (WFST/beam-search). פתרון זה יכול להגיע במקום או בנוסף לפתרונות המוצעים כיום, הכוללים את היכולת לאלץ על אוצר מילים מותאם אישית, או עדכון בסיסי של

המודל הקיים [63–64]. גישה זו תאפשר לסטארט-אפים קטנים להצטיין בדומיין מסויים על ידי אינטגרציה למודל אקוסטי נתון, במקום לאמן את המודל בעצמם, אופרציה יקרה מבחינת הון אנושי וידע בתחום. מאידך, חברות טכנולוגיה גדולות ייהנו מאימוץ רחב יותר של המודלים בתשלום שלהן.

כיצד ASR משתלב בנוף הרחב יותר של למידת מכונה?

מצד אחד, ASR הוא שווה בין שווים לראייה ממוחשבת (CV) ו-NLP אם התמלול זה המשימה הסופית. זהו המצב הנוכחי בשפות בהן יש אתגר באיסוף דאטה לאימון, ובדומיינים בהם התמלול הוא העסק העיקרי, למשל, בית משפט, רשומות רפואיות, כתוביות לסרטים וכו'.

מצד שני, ASR אינו עוד צוואר הבקבוק בתחומים בהם נפרץ סף שימושיות מסוים. במקרים אלה, ה-NLP הוא צוואר הבקבוק, מה שאומר ששיפור ביצועי ASR לקראת פרפקציוניזם אינו חיוני להפקת תובנות למשימה הסופית. לדוגמה, ניתן לייצר סיכום פגישה או לחלף את נקודות העניין משיחה באנגלית כבר כיום, על אף שהתמלול לא מושלם.

הערות לסיכום

ההתקדמות בטכנולוגיית ASR קירבה אותנו לעבר המטרה של תקשורת ללא חיכוך בין בני אדם למכונות, למשל ב-Conversational AI ו-Generative AI. עם המשך הפיתוח של מערכות לשיפור דיבור-ASR-דיאריזציה והופעתה של textless-NLP, אנו צפויים וצפויות לראות פריצות דרך מרגשות בתחום זה. בעודנו מצפים לעתיד, איננו יכולים שלא לצפות את האפשרויות האינסופיות שטכנולוגיית ASR תפתח.

תודה שהקדשת מזמנך לקרוא את הפוסט הזה! המחשבות והמשוב שלך על תחזיות אלה מוערכים מאוד. תגובות יתקבלו בברכה

הערות:

[1] <https://www.orcam.com/en/home/>

[2] <https://voicebot.ai/2022/12/01/hey-disney-custom-alexa-assistant-rolls-out-at-disney-world/>

[3] Jose, Christin, et al. "Latency Control for Keyword Spotting." ArXiv, 2022, <https://doi.org/10.21437/Interspeech.2022-10608>.

[4] Bijwadia, Shaan, et al. "Unified End-to-End Speech Recognition and Endpointing for Fast and Efficient Speech Systems." ArXiv, 2022, <https://doi.org/10.1109/SLT54892.2023.10022338>.

- [5] Yoon, Ji, et al. "HuBERT-EE: Early Exiting HuBERT for Efficient Speech Recognition." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2204.06328>.
- [6] Kanda, Naoyuki, et al. "Transcribe-to-Diarize: Neural Speaker Diarization for Unlimited Number of Speakers Using End-to-End Speaker-Attributed ASR." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2110.03151>.
- [7] Kanda, Naoyuki, et al. "Streaming Speaker-Attributed ASR with Token-Level Speaker Embeddings." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2203.16685>.
- [8] <https://www.fiercevideo.com/video/video-will-account-for-82-all-internet-traffic-by-2022-cisco-says>
- [9] Chiu, Chung, et al. "Self-Supervised Learning with Random-Projection Quantizer for Speech Recognition." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2202.01855>.
- [10] Chen, Sanyuan, et al. "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing." ArXiv, 2021, <https://doi.org/10.1109/JSTSP.2022.3188113>.
- [11] Hsu, Wei, et al. "Robust Wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2104.01027>.
- [12] Lugosch, Loren, et al. "Pseudo-Labeling for Massively Multilingual Speech Recognition." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2111.00161>.
- [13] Berrebbi, Dan, et al. "Continuous Pseudo-Labeling from the Start." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2210.08711>.
- [14] Radford, Alec, et al. "Robust Speech Recognition via Large-Scale Weak Supervision." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2212.04356>.
- [15] Graves, Alex, et al. "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks." ICML, 2016, https://www.cs.toronto.edu/~graves/icml_2006.pdf

- [16] Graves, Alex. "Sequence Transduction with Recurrent Neural Networks." ArXiv, 2012, <https://doi.org/10.48550/arXiv.1211.3711>.
- [17] Chan, William, et al. "Listen, Attend and Spell." ArXiv, 2015, <https://doi.org/10.48550/arXiv.1508.01211>.
- [18] Variani, Ehsan, et al. "Hybrid Autoregressive Transducer (Hat)." ArXiv, 2020, <https://doi.org/10.48550/arXiv.2003.07705>.
- [19] Meng, Zhong, et al. "Modular Hybrid Autoregressive Transducer." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2210.17049>.
- [20] Kim, Suyoun, et al. "Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2110.05376>.
- [21] Gulati, Anmol, et al. "Conformer: Convolution-Augmented Transformer for Speech Recognition." ArXiv, 2020, <https://doi.org/10.48550/arXiv.2005.08100>.
- [22] Vaswani, Ashish, et al. "Attention Is All You Need." ArXiv, 2017, <https://doi.org/10.48550/arXiv.1706.03762>.
- [23] Baevski, Alexei, et al. "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." ArXiv, 2020, <https://doi.org/10.48550/arXiv.2006.11477>.
- [24] Babu, Arun, et al. "XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2111.09296>.
- [25] <https://blog.google/technology/ai/ways-ai-is-scaling-helpful/>
- [26] <https://ai.facebook.com/blog/teaching-ai-to-translate-100s-of-spoken-and-written-languages-in-real-time/>
- [27] Ao, Junyi, et al. "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2110.07205>.

- [28] <https://github.com/guillaumekln/faster-whisper>
- [29] <https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/SpeechRecognition/wav2vec2>
- [30] Meng, Zhong, et al. "JEIT: Joint End-to-End Model and Internal Language Model Training for Speech Recognition." ArXiv, 2023, <https://doi.org/10.48550/arXiv.2302.08583>.
- [31] Zhang, Yu, et al. "Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages." ArXiv, 2023, <https://doi.org/10.48550/arXiv.2303.01037>.
- [32] Kendall, T. and Farrington, C. "The corpus of regional african american language". Version 2021.07. Eugene, OR: The Online Resources for African American Language Project. <http://oraal.uoregon.edu/coraal>, 2021
- [33] Brian, D Tran, et al. "Mm-hm," "Uh-uh": are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology?,' Journal of the American Medical Informatics Association, 2023, <https://doi.org/10.1093/jamia/ocad001>
- [34] Zhu, Ge, et al. "Filler Word Detection and Classification: A Dataset and Benchmark." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2203.15135>.
- [35] Anwar, Mohamed, et al. "MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation." ArXiv, 2023, <https://doi.org/10.48550/arXiv.2303.00628>.
- [36] Jaegle, Andrew, et al. "Perceiver IO: A General Architecture for Structured Inputs & Outputs." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2107.14795>.
- [37] <https://www.assemblyai.com/blog/conformer-1/>
- [38] Peyser, Cal, et al. "Dual Learning for Large Vocabulary On-Device ASR." ArXiv, 2023, <https://doi.org/10.48550/arXiv.2301.04327>.

- [39] Lakhotia, Kushal, et al. "Generative Spoken Language Modeling from Raw Audio." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2102.01192>.
- [40] Zeghidour, Neil, et al. "SoundStream: An End-to-End Neural Audio Codec." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2107.03312>.
- [41] Borsos, Zalán, et al. "AudioLM: a Language Modeling Approach to Audio Generation." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2209.03143>.
- [42] <https://about.fb.com/news/2022/10/hokkien-ai-speech-translation/>
- [43] Raffel, Colin, et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." ArXiv, 2019, /abs/1910.10683.
- [44] Sanh, Victor, et al. "Multitask Prompted Training Enables Zero-Shot Task Generalization." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2110.08207>.
- [45] Wang, Chengyi, et al. "Neural Codec Language Models Are Zero-Shot Text to Speech Synthesizers." ArXiv, 2023, <https://doi.org/10.48550/arXiv.2301.02111>.
- [46] Zhang, Ziqiang, et al. "Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling." ArXiv, 2023, <https://doi.org/10.48550/arXiv.2303.03926>.
- [47] <https://voicebot.ai/2022/07/05/eu-passes-new-regulations-for-voice-ai-and-digital-technology/>
- [48] <https://www.speechtechmag.com/Articles/ReadArticle.aspx?ArticleID=154094>
- [49] Thoppilan, Romal, et al. "LaMDA: Language Models for Dialog Applications." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2201.08239>.
- [50] Shuster, Kurt, et al. "BlenderBot 3: a Deployed Conversational Agent that Continually Learns to Responsibly Engage." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2208.03188>.

- [51] Xiaozhou, Zhou, et al. "Phonetic Embedding for ASR Robustness in Entity Resolution." Proc. Interspeech 2022, 3268–3272, doi: 10.21437/Interspeech.2022–10956
- [52] Chen, Angelica, et al. "Teaching BERT to Wait: Balancing Accuracy and Latency for Streaming Disfluency Detection." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2205.00620>.
- [53] Li, Qiujia, et al. "Improving Confidence Estimation on Out-of-Domain Data for End-to-End Speech Recognition." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2110.03327>.
- [54] <https://loora.ai/>
- [55] <https://techcrunch.com/2022/11/17/speak-lands-investment-from-openai-to-expand-its-language-learning-platform/>
- [56] Zhiqi, Huang, et al. "MTL-SLT: Multi-Task Learning for Spoken Language Tasks." NLP4ConvAI, 2022, <https://aclanthology.org/2022.nlp4convai-1.11>
- [57] Watanabe, Shinji, et al. "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings." ArXiv, 2020, <https://doi.org/10.48550/arXiv.2004.09249>.
- [58] Shafey, Laurent, et al. "Joint Speech Recognition and Speaker Diarization via Sequence Transduction." ArXiv, 2019, <https://doi.org/10.48550/arXiv.1907.05337>.
- [59] Kim, Juntae, and Lee, Jeehye. "Generalizing RNN-Transducer to Out-Domain Audio via Sparse Self-Attention Layers." ArXiv, 2021, <https://doi.org/10.48550/arXiv.2108.10752>.
- [60] Sathyendra, Kanthashree, et al. "Contextual Adapters for Personalized Speech Recognition in Neural Transducers." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2205.13660>.

[61] Tian, Jinchuan, et al. "Bayes Risk CTC: Controllable CTC Alignment in Sequence-to-Sequence Tasks." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2210.07499>.

[62] Tian, Zhengkun, et al. "Peak-First CTC: Reducing the Peak Latency of CTC Models by Applying Peak-First Regularization." ArXiv, 2022, <https://doi.org/10.48550/arXiv.2211.03284>.

[63] <https://docs.rev.ai/api/custom-vocabulary/>

[64] <https://cloud.google.com/speech-to-text/docs/adaptation-model>

Machine Learning

Speech Recognition

Generative Ai

Conversational Ai

Deep Dives