מהפכת הדיבור בעיצומה?

הודעה קולית או SMS?

עד כמה הטכנולוגיות באמת מנצלות את אחד ממקורות המידע הזמינים לנו ביותר - הדיבור? • האם ייתכן מצב שבו אדם שנפטר יקריא בקולו סיפור שנכתב 20 שנים לאחר מותו? מהו בעצם אמצעי התקשורת המועדף כיום בין אדם לחברו ובין אדם למכונה? **ד"ר ורד זילבר-ורוד**, חוקרת דיבור, בוחנת עד כמה הטכנולוגיות באמת מנצלות משהו שאנחנו עושים בו שימוש כל יום וכל היום - הדיבור

>> ורד זילבר-ורוד

כשם שדיבור הינו נדבך מרכזי בחיינו בהיבט של תקשורת בין אדם לאדם, כך דיבור יכול להיות חלק בלתי נפרד מחיינו בהיבט של תקשורת בין אדם למכונה. תקשורת שעד לאחרונה הייתה מושתתת בעיקר על טקסט כתוב ומודפס, מפות, למשל, נעזרת בדיבור כדי להנגיש את המידע - כמו במערכות ניווט קוליות במכונית.

בשנים האחרונות חלה התקדמות טכנולוגית גדולה שהביאה לגידול עצום בנפחי תעבורת המידע, שדיבור מהווה בו מרכיב חשוב. גידול זה גורם למהפך בצורך בשירותים וביישומים בכל תחום שבו קיימים תקשורת אדם-מכונה או תוכן מולטימדיה: החל מהאפשרות להכנסת קלט מדובר באופן מהיר - הכתבת מסרון (SMS), חיוג קולי בטלפון, חיפוש קולי במאגרי מידע - דרך האפשרות לשלוף מידע מתוך קטעי אודיו, וכלה באיתור מודיעיני מתוך דיבור בזמן אמת.

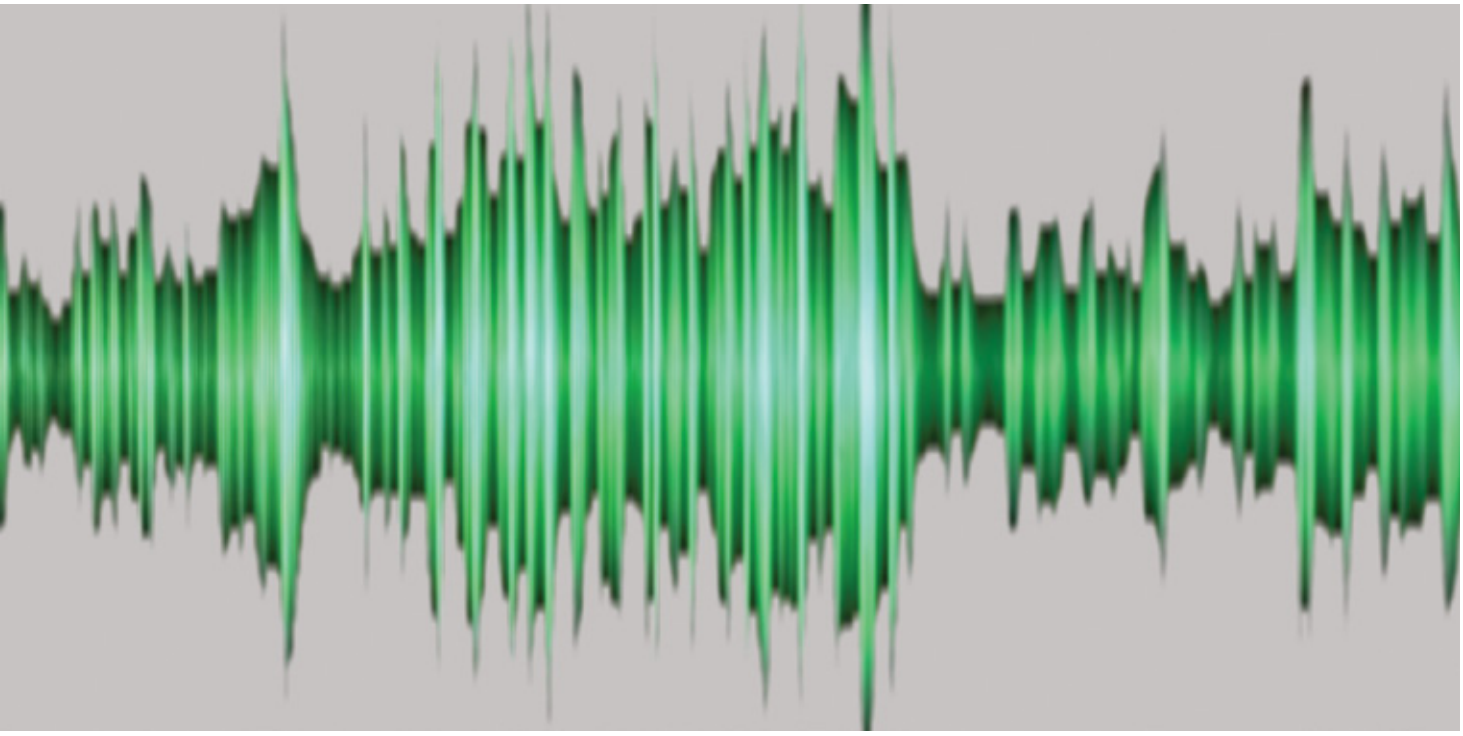
אתגרים ומשאבים

דיבור, בהשוואה לטקסט כתוב, הוא אות (signal) חולף ומגוון ביותר. שלא כמו כתב שנשאר על הנייר, מילה שכרגע נאמרה, ואינה מוקלטת, "חולפת עם הרוח". אם נדמיין לרגע דיאלוג בין בני אדם, הזמן שבו מבע מופק בפי הדובר ונתפס על ידי המאזין נקרא "זמן-אמת" (real-time) לעיבוד הדיבור.

מהנדסי טכנולוגיות הדיבור שואפים להדביק את קצב העיבוד הזה, כדי להפוך אותן לאטרקטיביות למשתמש. כל זה דורש פיתוח. ואולם, המכשול העיקרי בדרך לפיתוח טכנולוגיות דיבור בכל שפות העולם הוא

דיבור, בהשוואה לטקסט כתוב, הוא אות חולף ומגוון ביותר. שלא כמו כתב שנשאר על הנייר, מילה שכרגע נאמרה, ואינה מוקלטת "חולפת עם הרוח"

»



מתוך מגוון טכנולוגיות הדיבור הקיימות, יש תחום אחד שבו הטכנולוגיה מדביקה את היכולת האנושית והוא הפיכת טקסט לדיבור. כולנו זוכרים את המכשירים שהשמיעו דיבור בקול רובוטי

»

הצורך שלהן במשאבים רבים.

הקמת מנוע זיהוי דיבור קלאסי, למשל, מצריכה עבודת תשתית לא מבוטלת: הקלטה של דוברים רבים, המייצגים סטטיסטית את אוכלוסיית היעד, מבחינת מין, גיל, מבטא, וכל גורם חוץ-לשוני אחר; ייצוג של סביבות היעד - האם הזיהוי יתבצע רק בחדר שקט? ומדוע לא במכונית עם רעשי הרקע של כביש סואן? האם לקחת בחשבון את רעשי הרקע ברחוב? ומה עם סביבת בעלי חיים? ויש גם את ייצוג הערוצים - טלפון קווי בלבד? טלפון סלולארי? או רק מיקרופון בהרצאה?

אם לא די במגוון ובכמות ההקלטות (מאות שעות כדי להשיג ייצוג יחסי), יש להעביר אותן תהליך של תיוג ידני של תוכן כל משפט, כלומר, לתמלל אותן ידנית, לתת לתמלול ייצוג פונטי (ייצוג של רצף צלילי הדיבור בשפה), ולאחר מכן להעביר את בסיס הנתונים הזה תהליך אימון הדורש סיבוכיות חישובית גבוהה, ולכן גם זמן רב. יצוין, שכל תהליך אימון דורש ימים ואף שבועות ליצירת מודלים אקוסטיים ולשוניים.

המרחק להחלפת מאזין אנושי

עבודה סיזיפית זו, בעיקר בכל הקשור לתשתיות - איסוף של בסיסי הנתונים ותיוגם - שכאמור דורשת תקציב רחב היקף, עובדים שעברו הכשרה בתחום היחסית צר הזה, וטווח זמן של חודשים לאימון המודלים, מונעת את מיצוי הפוטנציאל הקיים בשימוש במערכות מבוססות זיהוי דיבור. המרחק עד לפריצת דרך, אשר תאפשר החלפת מאזין אנושי, עדיין רב.

ההתמודדות ההנדסית עם הגיוון הרב שהוצג לעיל מאתגרת מאוד.

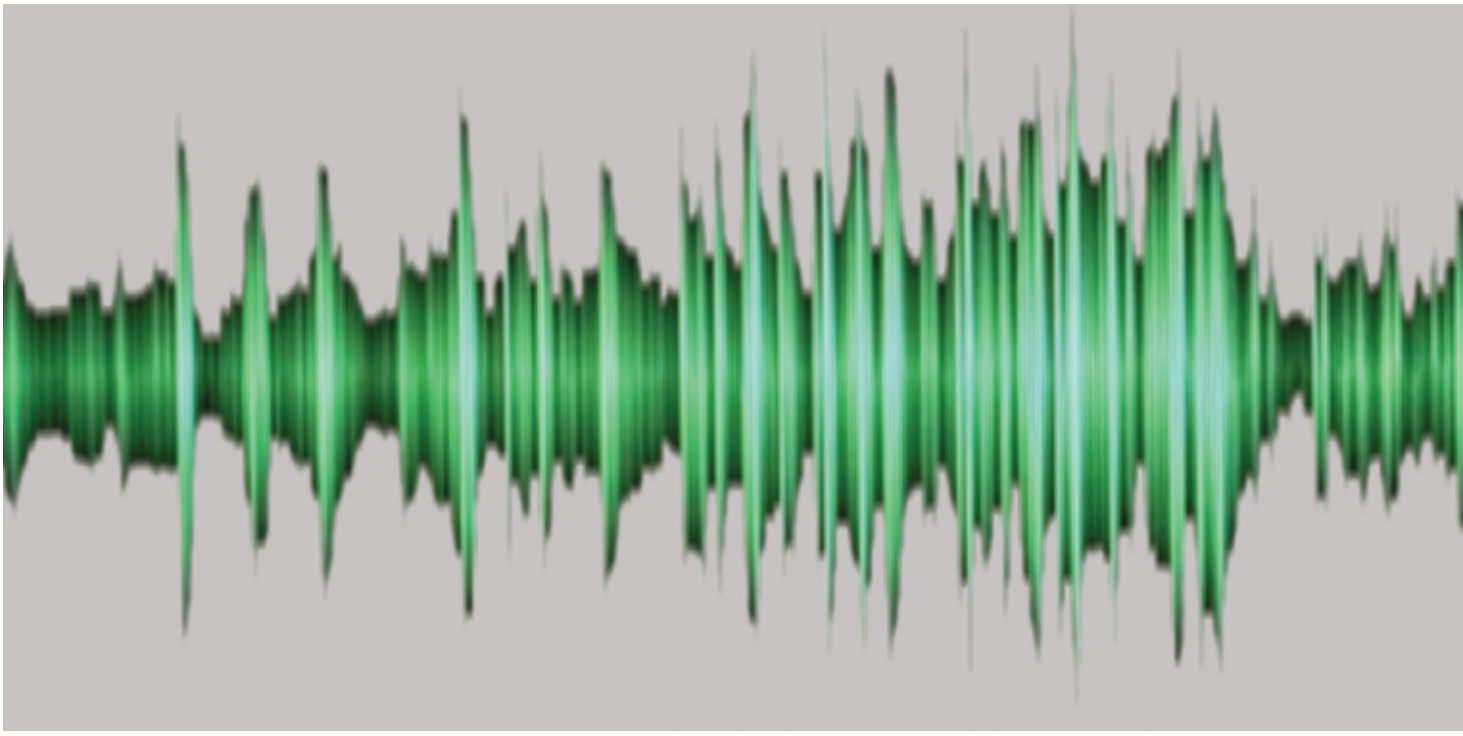
לצורך הדגמה מוצגות לפניכם תוצאות זיהוי דיבור במנוע הזיהוי של גוגל, המאפשר כיום חיפוש קולי בעברית, בשלושה סוגי דיבור: דיבור מוקרא, הרצאה בפני קהל ודיבור מתוך דיאלוג ספונטני. הקטעים המוצגים כוללים מבע בן 16 מילים. משך כל מבע הוא כ-10 שניות. בכל הקטעים מדובר בדוברת אישה, אך לא אותה אישה.

כאשר הקלט היה מתוך ספר מוקרא, הזיהוי היה כמעט מלא (המילים שזוהו באופן שגוי מודגשות):

- **טקסט מקורי מתוך ספר מוקרא (משך המבע 9.30 שניות):** "ספרים, עיתונים, מוסיקה, סרטים, תוכנות נתוני עסקות שבוצעו עם לקוחות, תוצאות משחקי כדורסל, שערי מניות בבורסה"
- **התמלול האוטומטי:** ספרים, עיתונים, מוזיקה, סרטים, תוכנות נתוני סקוטש בוצרים לקוחות, תוצאות משחקי כדורסל, שערי מניות בבורסה

כאשר הקלט היה מבע מתוך הרצאה בפני קהל, הזיהוי היה פחות טוב:

- **טקסט מקורי שתומלל במדויק וידנית (משך המבע 10.11 שניות):** "עכברים ארגונומיים, תחליפי עכבר, מערכות מיקוד מבט עבור אנשים שהם לקויי ראייה ועיוורים - יש לנו תוכנות"
- **התמלול האוטומטי:** עכברים ארגונומיים עכבר מערכות מיקוד מבט היא עבור אנשים עשירים, הביבים של תוכנות



הקוראים ישפטו את הזיהוי כאשר הקלט היה מבע מתוך דיבור ספונטני:

- **הטקסט המקורי שתומלל במדויק וידנית (משך המבע 5.68 שניות):** "בקשר לתרגום אני לא יודעת כאילו אמרתי לך במשך הסמסטר. אני לא יודעת כמה אני אספיק"
- **התמלול האוטומטי:** בקשר לתרגומים הולדת לחבר משחקים מטורנט

כדאי להשקיע בשמונה מיליון דוברי עברית?

מעבר למורכבות השפה האנושית, והגיוון במימוש השפה - כלומר הגיוון הרב שיש בדיבור (למשל, קול גבוה ונמוך, קול סדוק, דיבור איטי ומהיר), קיים אתגר החסינות לרעשים והתמודדות עם מגוון ערוצי התקשורת. לבסוף, פיתוח מנוע בשפה שבה קהל משתמשים הוא שמונה מיליון דוברים (עברית, למשל) פחות כדאי כלכלית למשקיעים, יחסית לפיתוח מנוע עבור שפות עתירות משאבים, כמו אנגלית, סינית, או ערבית. אולם, גם כאן, בקצב אטי אך מתמיד, החלחול לשפות דלות-משאבים מתרחש. זיהוי דיבור לשפה העברית הושק במנוע החיפוש של גוגל כבר בשנת 2011. חברת NUANCE, אחת מחברות הענק בתחום, השיקה לפני כשנה אפליקציית פיתוח לזיהוי השפה עברית, ומחלקות voice commands קיימות כיום בכל אחת מענקיות הטכנולוגיה (Google, Microsoft, Apple, Cisco, GM, ועוד).

קולה של אמא - גם כשהיא בקבר

מתוך מגוון טכנולוגיות הדיבור הקיימות, יש תחום אחד שבו הטכנולוגיה מדביקה את היכולת האנושית והוא - טקסט לדיבור (text-to-speech - TTS). כולנו זוכרים את המכשירים שהשמיעו דיבור בקול רובוטי. למערכות אלה הכינוי "דיבור סינטטי" אכן התאים.

כיום יש מערכות דיבור שאמנם מתבססות על מקטעי דיבור אמיתיים, של חלקיקי שניות, ושמסוגלות להרכיב מחדש ובצורה אוטומטית כל רצף של צלילי דיבור (פונמות) בשפה. במערכות כאלה קשה להבחין שלא מדובר באדם אלא במכונה, ורובן כבר אינן צורמות לאוזן האנושית.

עד כמה אנחנו באמת משתמשים בטכנולוגיית TTS? קרוב לוודאי שמלבד אתרים המחויבים לנגישות, לא פגשתם כמעט את השימוש בטכנולוגיה הזאת. אולם תארו לכם שלא רחוק היום, שבו תוכלו להשתמש בהקלטות של שעות דיבור ספורות של אמא שלכם (שנלקחו ממנה בעודה בחיים), ולשמוע את קולה מקריא לכם סיפור שנכתב 20 שנה אחרי מותה! נשמע דמיוני - אבל זה בהחלט יכול להתרחש במציאות!

כפי שתואר לעיל, אחד האגוזים הקשים ביותר לפיצוח בתחום טכנולוגיות דיבור הוא מערכות זיהוי דיבור, כלומר מערכות הממירות דיבור לטקסט כתוב (speech-to-text). הצורך בטכנולוגיה די ברור - להמיר כל אודיו לטקסט ולאחר מכן לעשות שימוש בטקסט לצורכי כריית מידע ממנו, יכול להאיץ את זרם האינפורמציה השוצף בערוצי התקשורת השונים.

כמו בכל תחום, גם טכנולוגיות הדיבור יפלו או יתרוממו כתוצאה מביקוש, כתוצאה מן הצורך שלנו, בני האדם. האם ההיסטוריה הלא-רחוקה של המשיבון הקולי, שהפך כיום למיותר, היא דוגמה לסיפור כולו?

»

ניצול הדיבור לתקשורת אדם-מכונה נמצא בנקודת זמן שבה הטכנולוגיות עדיין לא עומדות בדרישות האופטימאליות, אך נעשה בהן שימוש מושכל ביישומים מסוימים, בעיקר בשפות עתירות משאבים

כאשר מדובר בדוברים שהם אישים מפורסמים, כל ריאיון מתומלל כיום ידנית ומיד מתורגם לשפות העולם, אולם מה עם הרצאות אקדמיות? האם תמלול אוטומטי שלהן לא יאפשר חיסכון בזמן לסטודנטים?

"כמה טוב לשמוע את הקול שלך"

לדברי מומחים, הנאומים של הנשיא אובמה הפכו ללהיט בקרב לומדי האנגלית ביפן משום שהאנגלית שלו קלה להבנה. הוא מבטא מלים בצורה ברורה ומדבר בקצב אטי יחסית. סגנון הדיבור שלו מתאפיין בגילוי לב, הנקלט על ידי מאזיניו, ומשולב בצורת דיבור כמעט מוזיקלית.

במילים אחרות, בדומה לאינטראקציה בין דוברים, כדי ליצור ממשק יעיל של אדם-מכונה חייבים לקחת בחשבון את מרכיבי האינטונציה של הדיבור. מרכיבים אלה חיוניים לטכנולוגיות ממש כמו צלילי הדיבור (הפונמות), ויש להניח שנתונים פרוזודיים על הדיבור יכולים לייעל את עבודת המחשב.

בהקדמה לספרם של כהן, ג'יאנגולה ובאלו Voice User Interface Design (2004) נאמר שכבר בעיצוב המנוע יש לקחת בחשבון דפוסים פרוזודיים. ממש כפי שלעולם לא נכתוב מבע לא-דקדוקי כגון "בבקשה לדבר עכשיו אתה תל-אביב", כך אנחנו לא אמורים להפיק מבע שאינו דקדוקי מבחינה פרוזודית.

הקלטות של דיבור אינן דקדוקיות, כאשר המסרים הם, למשל, בעלי קונטור אינטונציה לא מתאים או זר; עם הפסקות במקומות לא מתאימים או בלי הפסקות היכן שצריך אותן; או עם הטעמות לקסיקליות ומקצביות שלא קיימות בשפה; או עם צלילים שאינם אפשריים מבחינה פיזיולוגית. על מנת להימנע ממצבים כאלה, מעצבי ממשק אדם-מכונה צריכים לשקף את הדפוסים ואת המבנים הרלוונטיים להקשרים מסוימים גם ברמה הפרוזודית שלהם, מה עוד שלכל שפה טבעית - ולכל דיאלקט של אותה שפה - יש את המבנה הדקדוקי הפרוזודי שלה.

פרוזודיה היא רכיב הכרחי בשפה הטבעית ממש כמו תחביר וסמנטיקה, ואין סיבה להניח שבזמן אינטראקציה עם מכונה (מחשב) הדובר יבטל את המרכיב הפרוזודי שבדיבור שלו. מעבר לכך, יש באינטונציה גם העברת רגש ודיוק של כוונות המוען. אכן, גם אם לא ביטאנו זאת ממש, האמירה "כמה טוב לשמוע את הקול שלך", היא מחשבה שכולנו מן הסתם חשנו בעבר. עובדה זאת שהופכת את הדיבור לבעל ערך מוסף על פני תקשורת בכתב.

המסרון הרג את המשיבון?

למרות כברת הדרך שעשתה הטכנולוגיה, מערכות המבוססות על זיהוי דיבור עדיין רחוקות מאוד מהפוטנציאל האמיתי הקיים בתחום זה, ואפילו במדינות עתירות משאבים, כמו ארה"ב, חלוצת המחקר של זיהוי דיבור, קיימים פערים, בעיקר טכנולוגיים, המונעים הנגשת השימוש לכלל המשתמשים, למשל: ילדים, דוברים מבוגרים מאוד ובעלי מבטאים זרים (כישראלים, כמעט כולנו חווינו כישלון במנועי חיפוש קוליים, כאשר דיברנו באנגלית למערכת הכתבת הודעות SMS בטלפון הסלולרי, ועוד). בישראל, ובמדינות בסדר גודל שלנו, עם שפה ייחודית ודלת-משאבים, קיימות אך ורק מערכות חלקיות שאינן מאפשרות שירותים מתקדמים.

ייתכן, שכמו בכל תחום, גם טכנולוגיות הדיבור יפלו או יתרוממו כתוצאה מביקוש, כתוצאה מן הצורך שלנו, בני האדם. האם ההיסטוריה הלא-רחוקה של המשיבון הקולי, שנתפש פעם כאביזר חיוני, אבל כיום הפך מיותר, היא דוגמה לסיפור כולו? הרי יותר ויותר אנשים מחליפים את המשיבון במסרונים ובמיילים, וחברות מסחריות הולכות לקראתם ומתרגמות את ההודעות הקוליות לטקסט.

ייתכן שהתהליך הזה מעיד על הכיוון שאליו נושבת רוח הטכנולוגיה - כל דיבור יהפוך לטקסט כתוב וכל טקסט כתוב יוכל להפוך לקולי. הכול כלול. המשתמש יחליט מה הוא לוקח מהבּוּפֶה. קרוב לוודאי שהמוען יעדיף להקריא את המסר, כאשר ידיו עסוקות במשהו אחר, ואילו הנמען יעדיף לרוב לקרוא את המסר. הטכנולוגיה תשרת אותנו בתווך.

האנושות תדלג על הדיבור?

התהליכים האלה חייבים להיות מלווים במחקר רב-דיסציפלינרי, מתוקף היותו מחקר על יחסי אדם-מכונה-אדם. בלשנים, מהנדסים, מדעני המחשב, חוקרי ההתנהגות האנושית ועוד, יכולים להתאגד לקהילת מחקר סינרגית בתחום טכנולוגיות הדיבור בארץ. יש לאפשר גישה פתוחה לתשתית משותפת לכל חברי הקהילה. דבר זה יאפשר פריצות דרך טכנולוגיות בתחום מאתגר זה. יתרון נוסף בקיום קהילה כזאת הוא בהיותה עמוד תווך בין האקדמיה לבין התעשייה שצמאה כיום לחוקרים מומחים בתחום.

אם לחזור לשאלה שבראש הדברים, ניצול הדיבור לתקשורת אדם-מכונה נמצא בנקודת זמן שבה הטכנולוגיות עדיין לא עומדות בדרישות האופטימאליות, אך נעשה בהן שימוש מושכל ביישומים מסוימים, ובעיקר בשפות עתירות משאבים.

אחת ממילות המפתח בעידן הצפת המידע היא שיתופיות, או שיתוף פעולה (collaboration), ובהקשר של תחום טכנולוגיות דיבור ייתכן שגישה כזאת תסייע לשפות שהן דלות משאבים להדביק את קצב הפיתוח של טכנולוגיות הדיבור. קשה להאמין שהאנושות תדלג על מקור המידע הזה ואמצעי התקשורת הפופולארי כל-כך - הדיבור.