התאוששות משיבושי טיסה

סוכן קולי יוצא בעברית לעדכון נוסעים ומתן החלטה מיידית

✓ נבחר

יוצא — עדכון יזום על שיבוש

כשטיסה מתעכבת או מתבטלת, הסוכן מתקשר יזום לכל נוסע, מעדכן ומציע טיסה חלופית או החזר כספי — בזמן אמת, ובמקביליות שאין למוקד אנושי.

נכנס — מוקד סטטוס קולי 24/7

הנוסע מתקשר ומזדהה, ומקבל מענה מיידי על סטטוס טיסה, כבודה ושינויי מושב — בלי המתנה בתור.

היוזקייס הנבחר: התאוששות משיבוש טיסה

זרימת השיחה

1. **זיהוי**
   אימות שמדובר בנוסע הנכון; טיפול במשיבון
2. **עדכון**
   מסירת פרטי השיבוש: טיסה, שעה מקורית, משך העיכוב
3. **הצגת אפשרויות**
   טיסה חלופית לעומת החזר כספי מלא
4. **קבלת החלטה**
   סיווג לתשובה, בקשת נציג, או בקשת חזרה מאוחר יותר
5. **סיכום וכתיבה**
   אישור בקול + כתיבת התוצאה חזרה למקור הנתונים

הבעיה

ביטול או עיכוב טיסה מייצר עשרות עד מאות שיחות נכנסות
- תוך דקות

גל השיחות מאריך זמני המתנה לכלל לקוחות המוקד, לא
- רק לנפגעי הטיסה

עיכוב בהחלטת הנוסע (חלופית / החזר) מעכב שחרור
- מושבים ומכירה מחדש

חוק שירותי תעופה (טיבי) מחייב מענה מהיר ומתועד לכל
- נוסע

ערך עסקי ו-ROI

בלימת גל שיחות

שיחה יזומה מונעת גל נכנס שמאריך המתנה לכלל הלקוחות

החלמה תפעולית מהירה

החלטות נסגרות מוקדם — שחרור ומכירה מחדש של מושבים

הפחתת חשיפה רגולטורית

מענה מתועד לכל נוסע לפי דרישות חוק שירותי תעופה

חוויית לקוח

עדכון אנושי ומהיר גם בשעות לא שגרתיות

מודל ROI לדוגמה — הנחות גלויות לאירוע שיבוש בודד

**ההנחות להלן הן דוגמה להמחשת מבנה החישוב, לא נתוני אל על בפועל*

רכיב	אנושי בלבד	(AI) משולב	חיסכון
עלות לאירוע שיבוש (200 שיחות)	~₪1,870	~₪930	**כ-50%**
שיעור הכלה (AI סוגר לבד, שמרני)	—	65%	—
משך ועלות שיחה	7 דק' · ~₪80/שעה	3 דק' · ~₪0.45/דק'	—

לא רק חיסכון — יכולת חדשה

200 שיחות בו-זמנית ב-3 לפנות בוקר הן בלתי אפשריות עבור מוקד אנושי. הערך האמיתי הוא לא רק % חיסכון — אלא היכולת עצמה.

השוואת ארכיטקטורות ומודלים

נבחר ✓ Cascaded — STT → LLM → TTS

שליטה טקסטואלית מלאה בכל שלב, function calling בשל, ובחירת ספק מיטבי לכל רכיב. קריטי כשנדרש דטרמיניזם בנוסח ובמספרים.

Speech-to-Speech (Realtime)

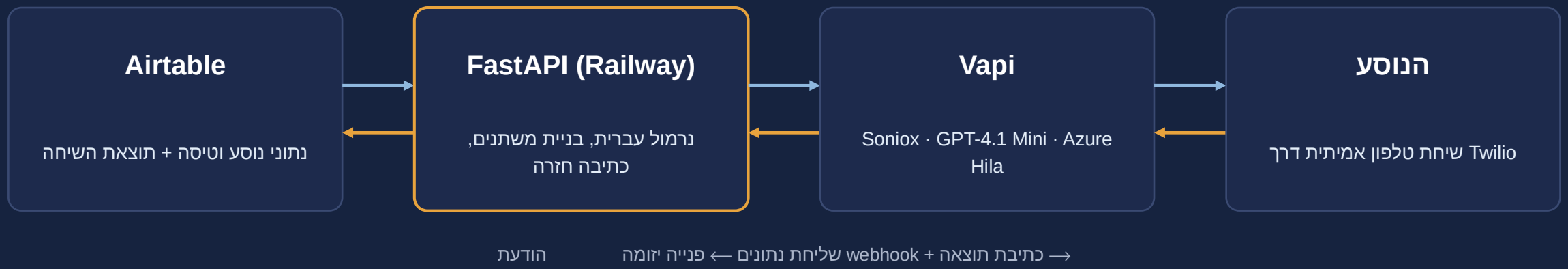
Latency נמוך יותר (300-600ms), אך שליטה חלקית בטקסט, כיסוי עברית לא מובטח, ו-function calling פחות בשל.

המודלים שנבחרו — לאחר בדיקה בפועל

רכיב	נבחר	החלופה שנבדקה	הנימוק לבחירה
STT	**Soniox (he)**	Deepgram Nova-3	WER 1.8% בעברית, 410ms, Keyterm Prompting למספרי טיסה
TTS	**Azure Hila**	ElevenLabs	עברית טבעית ויציבה; ElevenLabs Flash לא תומך בעברית כלל
LLM	**GPT-4.1 Mini**	GPT-4.1	עברית תקינה בבדיקה + חצי מהעלות ו-90ms פחות latency

ארכיטקטורת המערכת

Airtable · FastAPI (Railway) · Vapi — זרימה דו-כיוונית



נקודות מפתח בזרימה

3 מסלולים

GET לבדיקת נתונים, POST להפעלת שיחה, POST לקבלת webhook

אבטחה

אימות חתימת webhook (x-vapi-secret) + סודות רק ב-Environment Variables

חילוץ החלטה

ניתוח תמליל מקומי — לא תלוי בזמינות Structured Output של Vapi

עמידות

Retry עם exponential backoff, ותשובת 200 תמיד ל-Vapi כדי למנוע כפילויות

תוצאות ותובנות

שלוש שיחות יוצאות אמיתיות — הרשומות והוקלטו

נוסע	טיסה	החלטת הנוסע	סטטוס
יונתן כהן	LY315	החזר כספי מלא	**✓ הושלמה**
רוני לוי	LY315	טיסה חלופית	**✓ הושלמה**
דוד אברהם	LY315	בקשת נציג אנושי	**✓ הושלמה**

כל שיחה: זוהתה, עודכנה, קיבלה החלטה, וסוכמה חזרה למקור הנתונים אוטומטית — כולל טיפול נכון בבקשת נציג.

מגבלות שגילינו ומה הייתי משפר

הגייה לא עקבית של ספרות

Azure TTS קרא שעות וספרות בצורה לא טבעית — נבנתה שכבת נרמול לעברית מדוברת ("רבע לחצות" ולא "23:45")

Structured Output א-סינכרוני

נתוני ניתוח של Vapi הגיעו באיחור/לא הגיעו כלל — הוחלף בניתוח תמליל מקומי הזמין מיידית

התאמת ה"א" הידיעה בעברית

התאמת ביטויים דרשה נרמול של תחיליות ("הטיסה החלופית" מול "טיסה חלופית")

לשיפור עתידי

טיפול בשלילה בזיהוי כוונה ("לא טיסה חלופית") ומעבר לזיהוי מבוסס-LLM עבור שיחות מורכבות יותר