סיווג אודיו של שיחות טלפון — דו"ח ניסויים

1. מטרה

אימון מודל למידת מכונה שמסווג קטעי אודיו של שיחות טלפון לאחת מארבע קטגוריות:

קטגוריה	תיאור
`human`	נציג אנושי חי
`ivr`	תפריט אינטראקטיבי ("הקש 1 לעברית")
`music`	מוזיקת המתנה
`recording`	הקלטה / תא קולי / הודעה מוקלטת

המודל נדרש לתמוך גם בקטעים קצרים מ-5 שניות (עד 0.3 שניות), כי נציגים אנושיים עונים לעיתים במילה בודדת ("היי", "שלום").

2. מאגר הנתונים

הניסויים נערכו על שתי גרסאות של המאגר:

מאגר א' (ראשוני)

229 חתיכות לאחר חיתוך ל-5 שניות:

קטגוריה	חתיכות
human	103
ivr	54
music	60
recording	**12**

פיצול: אימון 158 / אימות 35 / בדיקה 36.

מאגר ב' (מורחב)

274 חתיכות — לאחר הוספת הקלטות `recording`:

קטגוריה	חתיכות
human	103
ivr	54
music	60

קטגוריה	חתיכות
recording	**57**

פיצול: אימון 189 / אימות 42 / בדיקה 43.

הערה מתודולוגית: הפיצול ל-train/val/test נעשה **לפי הקלטת מקור** (כל החתיכות מאותו קובץ הולכות לאותו פיצול) ועם stratification על הקטגוריה — כדי למנוע דליפת מידע (data leakage) שהייתה מנפחת את ציון הבדיקה.

3. ארכיטקטורת המודל

```
אודיו גולמי (16kHz, mono)
      ↓
wav2vec2-xls-r-300m  (encoder 300M, מאומן מראש, רב-לשוני — מוקפא (פרמטרים)
      ↓
Attention Pooling  (לומד אילו פריימים חשובים יותר)
      ↓
[concat: 15 פיצ'רים אקוסטיים מותאמי-קטגוריה]
      ↓
Linear(1024+15, 128) → GELU → Dropout(0.2) → Linear(128, 4)
      ↓
4 logits → Softmax → תחזית
```

- ה-encoder מוקפא; מתאמן רק ראש הסיווג (~133K פרמטרים מתוך 315M).
- Class weights מפצים על חוסר איזון בין הקטגוריות.
- Augmentations (רעש, time-stretch, pitch-shift, gain) על קבוצת האימון בלבד.
- Early stopping מבוסס macro F1 על קבוצת האימות.

4. הפיצ'רים האקוסטיים (Feature Engineering)

15 פיצ'רים, כל אחד מכוון להבחנה של קטגוריה ומבוסס על אלגוריתם מקובל (librosa). כולם מנורמלים לסקלה פיזית קבועה ונחתכים לטווח `[-5, 5]`.

קטגוריה יעד	פיצ'ר	אלגוריתם	הצדקה
music	`tempo_bpm`	beat tracking	מוזיקה בעלת קצב יציב
music	`beat_strength`	onset envelope	onset חזקים וסדירים
music	`harmonic_ratio`	HPSS	ספקטרום הרמוני נקי
music	`chroma_std`	chroma STFT	מגוון צלילים רחב
ivr	`dtmf_energy`	יחס אנרגיה בתדרי DTMF	צלילי חיוג

קטגוריה יעד	פיצ'ר	אלגוריתם	הצדקה
ivr	`pitch_stability`	std של pitch	קול סינתטי יציב
ivr	`silence_ratio`	יחס פריימים שקטים	פאוזות מובנות
human	`pitch_range`	אחוזון 5−95 של pitch	טון דינמי
human	`breath_events`	זיהוי נשימות	נשימות אנושיות
human	`voiced_ratio`	יחס פריימים קוליים	דיבור חי
recording	`bandwidth_hz`	rolloff 99%	פסקול דחוס (codec)
recording	`rolloff_hz`	rolloff 85%	רולוף נמוך
recording	`noise_floor_db`	אחוזון 10 של אנרגיה	רצפת רעש קבועה
כללי	`zcr_mean`	zero-crossing rate	הבחנה רעש/דיבור/מוזיקה
כללי	`spectral_centroid_hz`	spectral centroid	בהירות ספקטרלית

5. הניסויים

הגדרות קבועות בכל הריצות: מודל בסיס `wav2vec2-xls-r-300m`, batch size 8.

טבלת כל הריצות

#	מאגר	פיצ'רים	LR	unfreeze	patience	epochs	Test Acc	Macro F1
1	א' (rec=12)	5 נשימה (לא מנורמל)	1e-4	2	5	14	0.3889	0.2303
2	א' (rec=12)	5 נשימה (מנורמל)	1e-4	2	5	7	0.2500	0.1000
3	א' (rec=12)	**ללא**	5e-4	0	15	33	**0.9167**	**0.7094**
4	א' (rec=12)	5 נשימה (מנורמל)	5e-4	0	15	40	0.7500	0.5577
5	ב' (rec=57)	15 מותאמי-קטגוריה	1e-4	2	5	11	0.4186	0.3440
6	ב' (rec=57)	**15 מותאמי-קטגוריה**	5e-4	0	15	37	**0.7674**	**0.7518**
7	ב' (rec=57)	ללא	5e-4	0	15	21	0.5581	0.5426

F1 לכל קטגוריה

#	פיצ'רים	human	ivr	music	recording
1	5 נשימה (לא מנורמל)	0.4348	0.0000	0.4865	0.0000
2	5 נשימה (מנורמל)	0.0000	0.0000	0.4000	0.0000
3	ללא	0.9375	0.9000	1.0000	0.0000
4	5 נשימה (מנורמל)	0.7692	0.4615	1.0000	0.0000
5	15 (הגדרות גרועות)	0.4615	0.0000	0.5143	0.4000
6	**15 (הגדרות נכונות)**	0.7692	0.5714	1.0000	0.6667
7	ללא	0.0000	0.5294	0.9412	0.7000

6. ניתוח

6.1 ההגדרות חשובות יותר מהמודל

ריצות 1–2 (LR=1e-4, unfreeze=2, patience=5) קרסו: ה-`train_loss` כמעט לא ירד (1.39 → 1.29), כלומר המודל לא הצליח ללמוד אפילו את קבוצת האימון. עצם המעבר ל-LR=5e-4, הקפאת ה-encoder המלאה (unfreeze=0) ו-patience=15 (ריצה 3) הקפיץ את ה-Macro F1 מ-0.10 ל-0.71 — **על אותו מודל ואותו דאטה**. מסקנה: על מאגר קטן, fine-tuning של שכבות encoder מזיק (יותר מדי פרמטרים), ו-LR נמוך מדי מונע מראש הסיווג להתכנס.

6.2 ערך ה-Feature Engineering תלוי בדאטה

ממצא מרכזי ולא-טריוויאלי:

מאגר	בלי פיצ'רים	עם פיצ'רים
א' (rec=12)	**0.7094** (ריצה 3)	0.5577 (ריצה 4, 5 נשימה)
ב' (rec=57)	0.5426 (ריצה 7)	**0.7518** (ריצה 6, 15 מותאמים)

- **על מאגר א' הקטן והלא-מאוזן, הפיצ'רים הזיקו** — wav2vec2 לבד הספיק, והפיצ'רים (5 פיצ'רי נשימה, כולם על "דיבור אנושי") רק בלבלו את ההבחנה בין `ivr` ל-`human`.
- **על מאגר ב' המורחב, הפיצ'רים הפכו לקריטיים** — בלעדיהם המודל **קרס לחלוטין על `human`** (F1=0.00: כל 15 ההקלטות האנושיות סווגו כ-ivr או recording). הפיצ'רים העלו את human מ-0.00 ל-0.77.

ההסבר: ככל שהמאגר גדל והתאזן (recording 12→57), המשימה נעשתה קשה יותר — ארבע קטגוריות שכולן (חוץ ממוזיקה) מכילות דיבור. במצב הזה, פיצ'רים גלובליים מותאמי-קטגוריה (tempo, DTMF, רוחב-פס, דינמיקת pitch) מספקים אות משלים ש-wav2vec2 — המאומן על דיבור — לא תופס לבדו.

6.3 השפעת הוספת הדאטה ל-recording

בכל הריצות על מאגר א' (recording=12), קטגוריית `recording` קיבלה 8 — F1=0.00 דוגמאות אימון פשוט לא הספיקו. לאחר ההרחבה ל-57, `recording` הגיע ל-F1=0.67–0.70. זו הדגמה ישירה לחשיבות כמות הדוגמאות לכל מחלקה.

7. מטריצות בלבול (ריצות מרכזיות, מאגר ב')

ריצה 6 — 15 פיצ'רים + הגדרות נכונות (המודל הנבחר):

```
                חיזוי
אמת        human  ivr  music  recording
human        15    0     0      0
ivr           5    4     0      1
music         0    0     9      0
recording     4    0     0      5
```

ריצה 7 — בלי פיצ'רים:

```
                חיזוי
אמת        human  ivr  music  recording
human         0   12     0      3      ← human קריסה מוחלטת על
ivr           0    9     0      1
music         0    1     8      0
recording     0    2     0      7
```

8. מסקנות

1. **המודל הנבחר: ריצה 6** — wav2vec2-xls-r-300m מוקפא + 15 + attention pooling פיצ'רים אקוסטיים, על המאגר המורחב. **דיוק 76.7%, Macro F1 = 0.75.**
2. **Feature engineering תורם** על המאגר הנוכחי: +0.21 ב-Macro F1 לעומת המודל בלי פיצ'רים, והצלת קטגוריית `human` מקריסה.
3. **כיוונון hyperparameters קריטי**: אותם פיצ'רים עם הגדרות שגויות נתנו רק 0.34.
4. **כמות הדאטה לכל מחלקה מכריעה**: `recording` עבר מ-F1=0 (8 דוגמאות) ל-F1=0.67 (39 דוגמאות).

9. מגבלות וכיווני המשך

- **`ivr` עדיין החלש** (F1=0.57, recall 0.40) — מתבלבל עם `human`. כיוונים: עוד דאטה ל-ivr, או חיזוק הפיצ'רים המכוונים אליו (dtmf_energy, pitch_stability).
- **מאגר קטן יחסית** (274 חתיכות) — ביצועים עשויים להשתפר עם עוד דאטה ובדיקה על מאגר חיצוני.
- **חוסר איזון שיורי** — recording עדיין הקטגוריה הקטנה ביותר.