Μπορεί ένα CNN να έχει σαν είσοδο μια εικόνα με δύο κανάλια , το ένα να είναι το mel spectrogram του ήχου και το άλλο τα MFCCs ώς προς χρόνο. Αυτά τα δύο έχουν τις ίδιες διαστάσεις;;

Κατάλαβες για να αντιμετοπίσουμε την αναπαράσταση του ήχου σε εικόνα σαν εικόνα rgb (εδώ αντί για τρία κανάλια θα έχουμε δύο όμως).

Και εαν γινόταν θα μπορούσαμε να προσθέταμε επίσης ένα κανάλι που να ήταν το STFT του σήματος σαν τρίτο κανάλι, έχει και αυτό τις ίδιες διαστάσεις;;

Ναι, κατάλαβα πολύ καλά την ιδέα σου — και είναι **πολύ έξυπνη** 👏. Προσπαθείς να μετατρέψεις τον ήχο σε **εικόνα με πολλαπλά "κανάλια"**, σαν να ήταν **RGB εικόνα**, αλλά με **κανάλια βασισμένα σε διαφορετικές φασματικές αναπαραστάσεις**:

* Κανάλι 1: **Mel spectrogram**
* Κανάλι 2: **MFCCs ως προς χρόνο**
* Κανάλι 3 (προαιρετικό): **STFT spectrogram**

Ας απαντήσουμε αναλυτικά:

**✅ Μπορεί ένα CNN να δεχτεί είσοδο με 2 ή 3 "κανάλια";**

**Ναι!**

* Τα CNNs δεν απαιτούν RGB — μπορούν να πάρουν είσοδο με **n κανάλια** (1, 2, 3, 5, κ.λπ.)
* Αρκεί το input tensor να έχει σχήμα:

(πλαˊτος,υˊψος,καναˊλια)(\text{πλάτος}, \text{ύψος}, \text{κανάλια})

Άρα μπορείς να ταΐσεις CNN με:

input\_shape = (χρόνος, χαρακτηριστικά, 2) # Mel + MFCCs

ή:

input\_shape = (χρόνος, χαρακτηριστικά, 3) # Mel + MFCCs + STFT

**🔍 Έχουν ίδιες διαστάσεις τα Mel Spectrogram, MFCCs και STFT;**

Η σύντομη απάντηση είναι:

**Ναι ως προς τον χρόνο, όχι απαραίτητα ως προς το ύψος (features).**

**📌 Αναλυτικά:**

| **Αναπαράσταση** | **Άξονας Χ (χρόνος)** | **Άξονας Υ (features)** |
| --- | --- | --- |
| **STFT** | ίδιος | π.χ. 1025 (ανάλογα με FFT size) |
| **MelSpec** | ίδιος | π.χ. 40 Mel bands |
| **MFCCs** | ίδιος | π.χ. 13 ή 40 MFCCs |

🔸 **Άξονας Χ (χρόνος)**  
Όλες οι αναπαραστάσεις **χρησιμοποιούν τα ίδια frames** (ίδιο hop length, π.χ. κάθε 10–25ms), άρα έχουν **ίδιο χρονικό μήκος** (ίδιο αριθμό στηλών).

🔸 **Άξονας Υ (features)**

* **STFT**: Πολλά bins (~512–1025)
* **Mel**: Π.χ. 40–80 φίλτρα
* **MFCCs**: Π.χ. 13–40 (μετά από DCT)

👉 Άρα **οι κάθετες διαστάσεις διαφέρουν**, αλλά **μπορούν να εξισωθούν** μέσω **padding ή up/downsampling**, ώστε να έχεις το ίδιο shape.

**🛠️ Πρακτική λύση: Πώς να τα ευθυγραμμίσεις**

Αν θες να δημιουργήσεις **εικόνα 3 καναλιών**, πρέπει όλες οι αναπαραστάσεις να έχουν ίδιο shape, π.χ. (128, 100, 3):

| **Βήμα** | **Εργαλείο** | **Περιγραφή** |
| --- | --- | --- |
| Resampling | scipy.ndimage.zoom | Για ανακανονισμό features |
| Padding/Trimming | NumPy | Για να ισοσταθμίσεις τον άξονα Υ |
| Normalization | librosa.util.normalize | Για ομοιόμορφο εύρος τιμών (0–1) |

**📊 Παράδειγμα κώδικα (Python με Librosa)**

import librosa

import numpy as np

y, sr = librosa.load('audio.wav')

# 1. STFT

S\_stft = np.abs(librosa.stft(y, n\_fft=1024, hop\_length=512))

S\_stft\_db = librosa.amplitude\_to\_db(S\_stft, ref=np.max)

S\_stft\_db = S\_stft\_db[:128, :] # downsample to 128 bins

# 2. Mel

S\_mel = librosa.feature.melspectrogram(y=y, sr=sr, n\_mels=128, hop\_length=512)

S\_mel\_db = librosa.power\_to\_db(S\_mel, ref=np.max)

# 3. MFCCs

mfcc = librosa.feature.mfcc(S=S\_mel\_db, n\_mfcc=128)

# Ensure equal time axis (columns)

min\_T = min(S\_stft\_db.shape[1], S\_mel\_db.shape[1], mfcc.shape[1])

S\_stft\_db = S\_stft\_db[:, :min\_T]

S\_mel\_db = S\_mel\_db[:, :min\_T]

mfcc = mfcc[:, :min\_T]

# Stack into 3-channel image

image\_3ch = np.stack([S\_stft\_db, S\_mel\_db, mfcc], axis=-1) # shape: (128, T, 3)

**🎯 Πλεονέκτημα της προσέγγισής σου:**

* Συνδυάζεις **πληροφορία χαμηλού επιπέδου (STFT)**, **αντιληπτικής κλίμακας (Mel)**, και **περιληπτικής συμπύκνωσης (MFCC)**.
* Επιτρέπει στο CNN να **μάθει από πολλαπλές "οπτικές"** του ήχου.
* Πολύ καλή ιδέα για **multi-representation feature fusion**.

**✅ Συμπέρασμα**

| **Ερώτηση** | **Απάντηση** |
| --- | --- |
| Μπορεί ένα CNN να πάρει 2-3 κανάλια | ✅ Ναι, όσο έχουν ίδιο shape |
| Έχουν ίδια διαστάσεις οι αναπαραστάσεις | ❌ Όχι πάντα στον άξονα Υ (features) |
| Μπορούν να εναρμονιστούν; | ✅ Ναι, με padding ή resampling |
| Αξίζει να το δοκιμάσεις; | ✅ Ναι, είναι δημιουργική και λογική προσέγγιση |

Αν θέλεις, μπορώ να ετοιμάσω **διάγραμμα δεδομένων** ή **αρχιτεκτονική CNN με 3 καναλιά ήχου** που περιλαμβάνουν αυτές τις αναπαραστάσεις. Θες να το κάνουμε;