Κατασκευή Δεδομένων για την Εκπαίδευση και Αξιολόγηση Συστημάτων Τεχνητής Νοημοσύνης στον Εντοπισμό Ηχητικών Συμβάντων

Κωνσταντίνος Θεόδωρος Τσάμης[1], Αιμίλιος Βασίλειος Καμπουρόπουλος[2], Μάξιμος Καλιακάτσος Παπακώστας[1]

[1]Τμήμα Μουσικής Τεχνολογίας & Ακουστικής, Σχολή Μουσικής και Οπτοακουστικών Τεχνολογιών, Ελληνικό Μεσογειακό Πανεπιστήμιο, Ε. Δασκαλάκη, Περιβόλια, 74133, Ρέθυμνο

[2]Τμήμα Μουσικών Σπουδών, Σχολή Καλών Τεχνών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Πανεπιστημιούπολη Θέρμης, 54124

tsamkonstheod@gmail.com

Περίληψη

Στην παρούσα εργασία παρουσιάζεται ένα σύστημα ανίχνευσης ηχητικών συμβάντων το οποίο εκπαιδεύεται με τη χρήση τεχνητού συνόλου δεδομένων που περιλαμβάνει μίξεις ήχων. Το μοντέλο βασίζεται στο Wav2Vec2, ένα προεκπαιδευμένο μοντέλο τεχνητής νοημοσύνης για την αναγνώριση ομιλίας, και σε ένα LSTM δίκτυο για να προβλέψει την ταυτόχρονη επικάλυψη ηχητικών γεγονότων. Τα δεδομένα εκπαιδεύονται με ηχητικά κανάλια που περιλαμβάνουν έως πέντε ταυτόχρονους ήχους. Τα πρώτα αποτελέσματα είναι ενθαρρυντικά, ωστόσο φαίνεται πως η ακρίβεια μειώνεται σε πολύπλοκες μίξεις με μεγαλύτερη πυκνότητα ήχων, γεγονός που υποδεικνύει τα περιθώρια βελτίωσης μελλοντικά.

Data Construction for Training and Evaluation of Artificial Intelligence Systems in Audio Event Detection

ABSTRACT

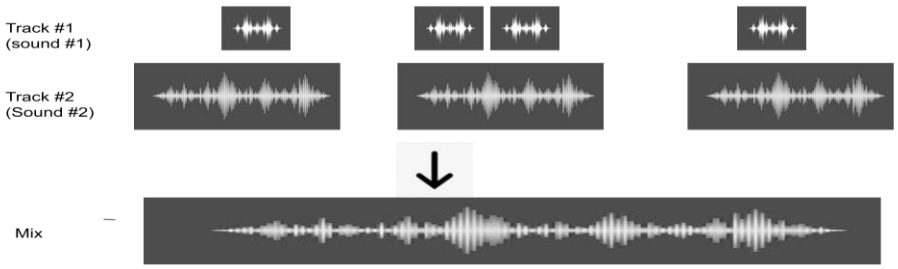
This paper presents a system for detecting audio events, which has been trained using an artificial dataset that includes sound mixtures. The model is based on Wav2Vec2, a pre-trained AI model for speech recognition, and uses LSTM to predict overlapping audio events. The data has been trained on audio channels that contain up to five simultaneous sounds. The initial results are promising, but accuracy decreases in more complex mixtures with higher sound density, indicating the potential for improvements in the future.

Εισαγωγή

Η ανίχνευση ηχητικών συμβάντων (event detection) αποτελεί ένα σημαντικό πεδίο έρευνας στην επεξεργασία ήχου με στόχο τον εντοπισμό και την καταμέτρηση γεγονότων και πηγών σε ηχητικά αρχεία. Η ανάγκη για τέτοιου είδους εργαλεία παρουσιάζεται σε διάφορους τομείς, όπως παραδείγματος χάρη η ανάλυση μουσικής [1], η επεξεργασία ήχου και βίντεο, η παρακολούθηση της άγριας ζωής [2], αλλά και η υγειονομική περίθαλψη [3]. Πρόσφατη πρόοδος στους αλγόριθμους μηχανικής μάθησης έχει φέρει ως αποτέλεσμα την ανάπτυξη μεθόδων που μπορούν να χαρακτηρίσουν ένα απομονωμένο ηχητικό απόσπασμα [4] ή να διαχωρίσουν πηγές από μίξη ηχητικού σήματος που περιλαμβάνει συγκεκριμένες ασθενώς επισημειωμένες πηγές [5], όμως η ανίχνευση και η καταμέτρηση διακριτών ηχητικών γεγονότων σε πολύπλοκα ηχητικά αρχεία με επικάλυψη πολλαπλών πηγών ήχου παραμένει μια πρόκληση. Στο παρόν κείμενο, αναλύεται η διαδικασία εκπαίδευσης ενός μοντέλου τεχνητής νοημοσύνης, με τη χρήση ενός αλγορίθμου δημιουργίας τεχνητών συνόλων δεδομένων (dataset), με στόχο την ανίχνευση αλληλεπικαλυπτόμενων ηχητικών συμβάντων σε ηχητικά αρχεία.

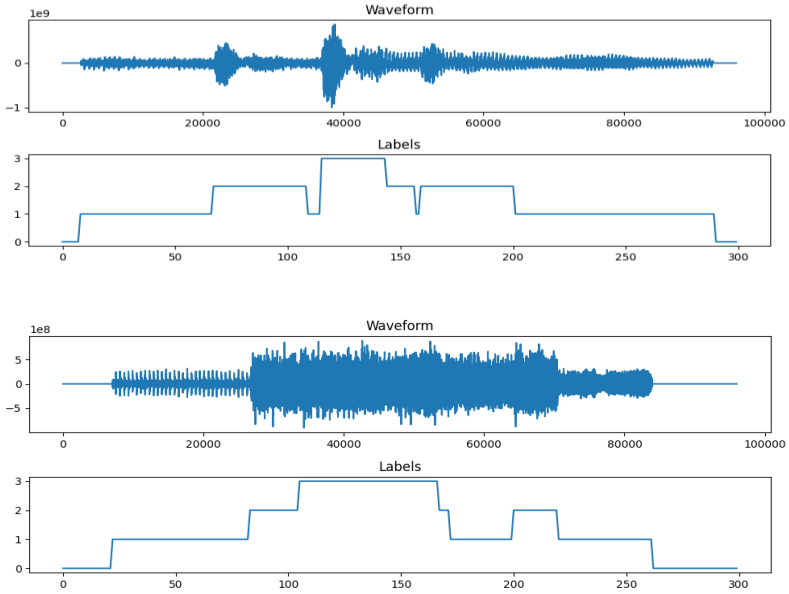
1. Μέθοδος

Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκε ένα τεχνικό dataset που δημιουργήθηκε μέσω μίξης διαφορετικών κομματιών ήχου. Η διαδικασία δημιουργίας του dataset υλοποιήθηκε με την ανάπτυξη μιας κλάσης, η οποία κατασκευάζει τυχαίες μίξεις ήχου και τις αντίστοιχες επισημάνσεις των ηχητικών συμβάντων. Οι πρωταρχικοί ήχοι είναι επιλεγμένοι από την ιστοσελίδα freesound.org [1] και εφαρμόζονται πάνω τους φίλτρα για τον εμπλουτισμό και μεγέθυνση του dataset. Συγκεκριμένα, μπορούν να εφαρμοστούν σε κάθε ήχο τα εξής φίλτρα: χαμηλοπερατό φίλτρο (lowpass filter), υψιπερατό φίλτρο (highpass filter), αλλαγή τονικού ήχους (pitch shift) και κοκκώδη σύνθεση (granular synthesis). Έπειτα οι ήχοι χρησιμοποιούνται από την κλάση που δημιουργήθηκε τοποθετώντας τυχαίες επιλογές από τους επεξεργασμένους ήχους σε κανάλια με τυχαία ένταση και τελικά δημιουργώντας μια μίξη. Κάθε κανάλι περιέχει μόνο έναν ήχο τοποθετημένο σε τυχαίες θέσεις και εντάσεις. Κατά την εκτέλεση του προγράμματος καθορίζονται οι παράμετροι: duration (διάρκεια), που ορίζει την διάρκεια της κάθε μίξης, και number_of_tracks (αριθμός καναλιών), που ορίζει πόσα κανάλια θα παραχθούν. Η επιλογή μόνο 1 καναλιού, θα δημιουργήσει μια μίξη χωρίς επικαλύψεις απαρτιζόμενη από έναν μόνο ήχο. Ενώ η επιλογή 5, για παράδειγμα, καναλιών, θα δημιουργήσει μια μίξη με 5 ήχους (1 σε κάθε κανάλι) και επικάλυψη το πολύ μέχρι 5 ταυτόχρονους ήχους.



Σχήμα 1: Διαδικασία παραγωγής μίξης ήχου

Για κάθε μίξη, παράγονται και οι αντίστοιχες επισημάνσεις, οι οποίες είναι απαραίτητες για την εκπαίδευση του μοντέλου. Οι επισημάνσεις αυτές είναι μια λίστα τιμών που αντιστοιχεί σε διαστήματα των 20 ms του ήχου και δηλώνει τον αριθμό των ήχων που επικαλύπτονται κάθε χρονική στιγμή.



Σχήμα 2: επισημειώσεις τεχνητού συνόλου δεδομένων

Το πρόγραμμα απαρτίζεται από το Wav2Vec2 μοντέλο και στην έξοδο του ένα LSTM δίκτυο.

Το Wav2Vec2 είναι ένα μοντέλο βαθιάς μάθησης που αναπτύχθηκε από την ομάδα της Meta AI (πρώην Facebook AI Research) για την αναγνώριση ομιλίας. Βασίζεται σε μια προσέγγιση αυτόματης μάθησης χαρακτηριστικών, με στόχο να εκπαιδευτεί πάνω σε μη επισημειωμένα δεδομένα και να προσφέρει αναπαραστάσεις για μοντέλα downstream, όπως η αναγνώριση φωνής και άλλες εργασίες που σχετίζονται με την επεξεργασία ήχου. Το μοντέλο χωρίζεται σε δύο κύρια μέρη, τον encoder, έναν αρχικό κωδικοποιητή (convolutional neural network - CNN) που λαμβάνει το ακατέργαστο σήμα και το μετατρέπει σε μια χρονική αναπαράσταση, διατηρώντας σημαντικές πληροφορίες σχετικά με το περιεχόμενο του ήχου σε μικρότερη κρυφή διάσταση (latent space) και τον contextualizer, ένα μετασχηματιστή (Transformer) που λαμβάνει την έξοδο του κωδικοποιητή και κατασκευάζει αναπαραστάσεις που λαμβάνουν υπόψη τις μακροπρόθεσμες εξαρτήσεις στο σήμα. Χρησιμοποιώντας self-attention, το μοντέλο μπορεί να κατανοήσει το πλαίσιο των ήχων σε βάθος χρόνου. Το Wav2Vec2 χρησιμοποιεί μη επιβλεπόμενη μάθηση (unsupervised learning), όπου μαθαίνει να προβλέπει τμήματα του σήματος ήχου ανά 20 ms που έχουν "καλυφθεί" (masked) κατά τη διάρκεια της εκπαίδευσης. Αυτή η τεχνική επιτρέπει στο μοντέλο να μαθαίνει ουσιαστικές αναπαραστάσεις χωρίς την ανάγκη για μεγάλα σύνολα δεδομένων με επισήμανση.

Το LSTM (Long Short-Term Memory) είναι ένας τύπος νευρωνικού δικτύου αναδρομής (Recurrent Neural Network - RNN) που σχεδιάστηκε ειδικά για να αντιμετωπίσει το πρόβλημα της μακροχρόνιας εξάρτησης (long-term dependency) στα παραδοσιακά RNNs. Το LSTM αποτελείται από την κυψέλη μνήμης (memory cells), τα οποία έχουν την ικανότητα να διατηρούν ή να απορρίπτουν πληροφορίες. Κάθε κύτταρο LSTM έχει τρεις βασικές πύλες, την πύλη εισόδου (input gate) που αποφασίζει ποιες νέες πληροφορίες θα προστεθούν στη μνήμη, την πύλη λήθης (forget gate), που καθορίζει ποιες πληροφορίες θα απορριφθούν ή θα ξεχαστούν από την κυψέλη μνήμης και την πύλη εξόδου (output gate) υπεύθυνη να ρυθμίζει το ποιο μέρος της μνήμης θα χρησιμοποιηθεί για να παραχθεί η έξοδος.

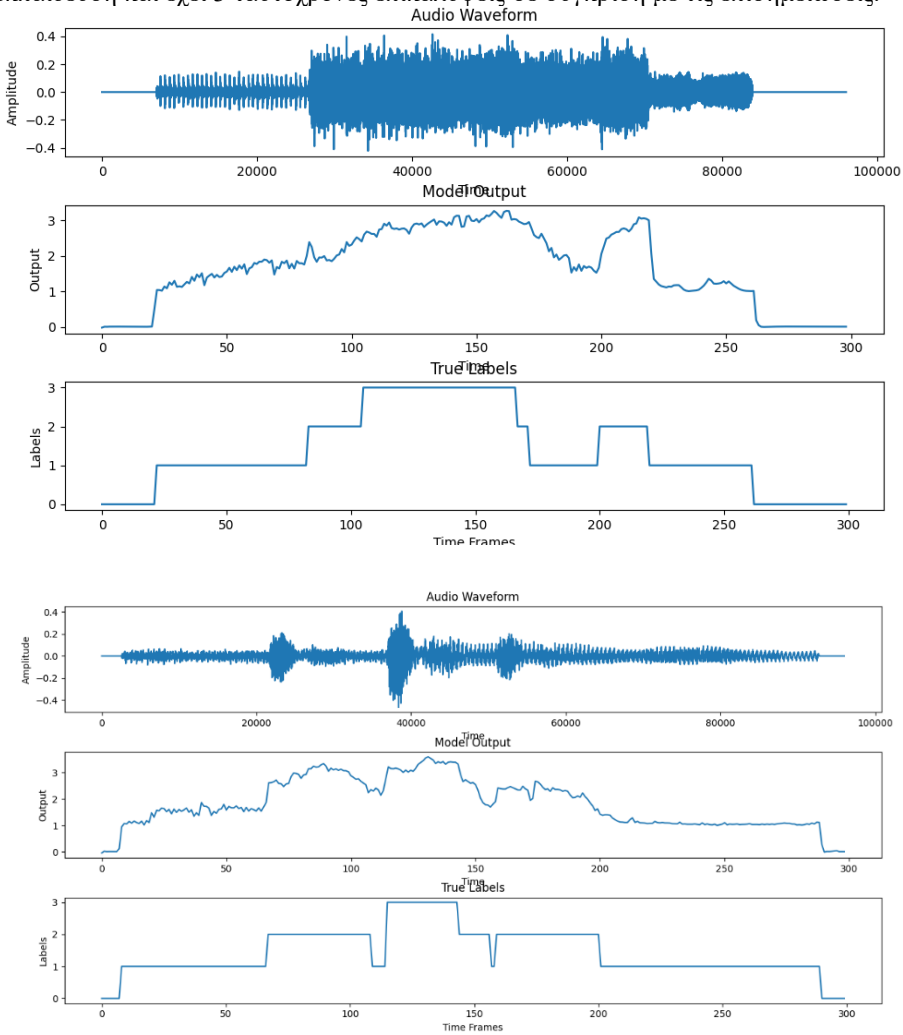
Αυτές οι πύλες επιτρέπουν στο LSTM να διατηρεί χρήσιμες πληροφορίες για μεγαλύτερες χρονικές ακολουθίες, κάτι που το καθιστά εξαιρετικά αποτελεσματικό για εργασίες που σχετίζονται με ακολουθίες, όπως η επεξεργασία ήχου, κειμένου, και βίντεο.

Πιο συγκεκριμένα, αξιοποιήθηκε το μοντέλο «facebook/wav2vec2-base» που έχει έξοδο μια αναπαράσταση 768 διαστάσεων. Στην συνέχεια το LSTM μειώνει τις διαστάσεις σε 256 και 128 και, εν τέλει, σε 1 ενώ ταυτόχρονα παρεμβάλει και μια μη-γραμμική συνάρτηση ενεργοποίησης τύπου ReLU.

2. Αποτελέσματα

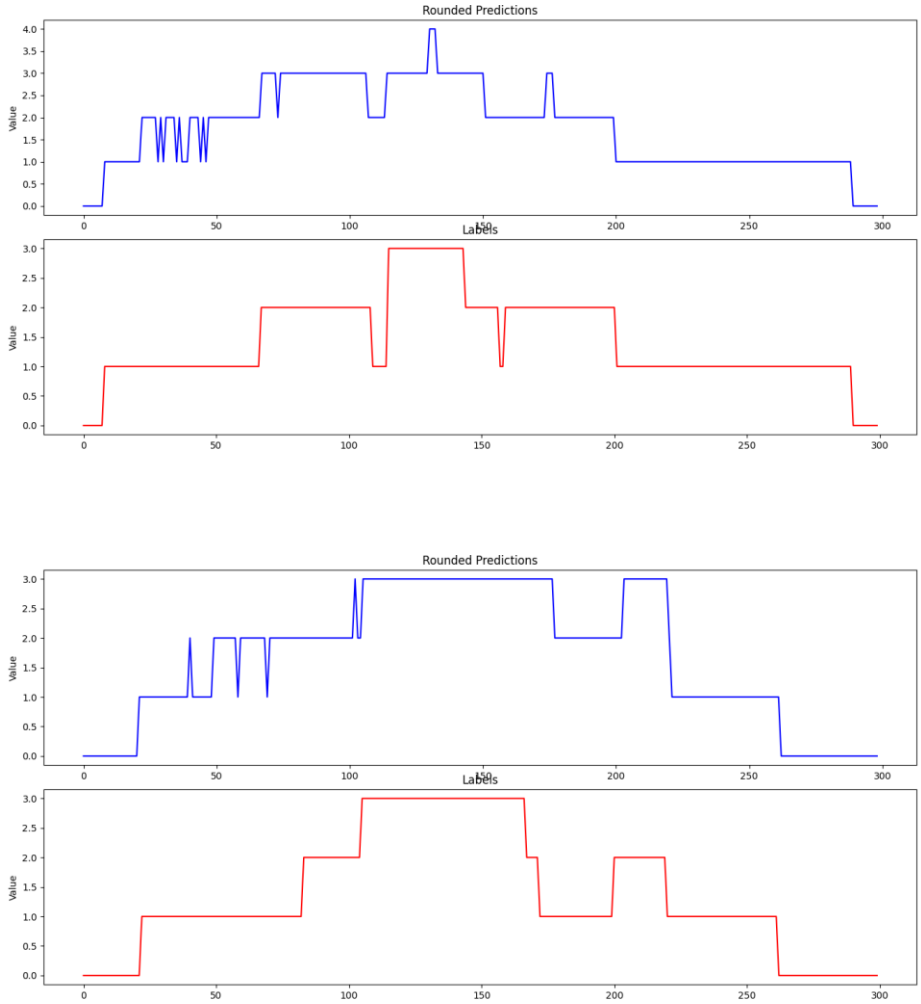
Έπειτα από την εκπαίδευση του μοντέλου σε αρχεία ήχου με μικρή πυκνότητα επικάλυψης το μοντέλο καταφέρνει να προβλέψει σωστά μεγάλο μέρος των απαντήσεων. Η δυσκολία του σε μεγαλύτερη πυκνότητα επικάλυψης ήχων οφείλεται

κυρίως στο γεγονός ότι κατά την εκπαίδευση, για λόγους εξοικονόμησης πόρων, εκπαιδεύτηκε με μέγιστο αριθμό καναλιών ίσο με 5. Συνεπώς, οι μέγιστες ταυτόχρονες επικαλύψεις μπορεί να είναι το πολύ 5. Στο σχήμα που ακολουθεί φαίνονται οι προβλέψεις του μοντέλου σε ένα ηχητικό αρχείο που δεν έχει χρησιμοποιηθεί κατά την εκπαίδευση και έχει 3 ταυτόχρονες επικαλύψεις σε σύγκριση με τις επισημειώσεις.



Σχήμα 3: Πρόβλεψη μοντέλου και σύγκρισή με επισημειώσεις

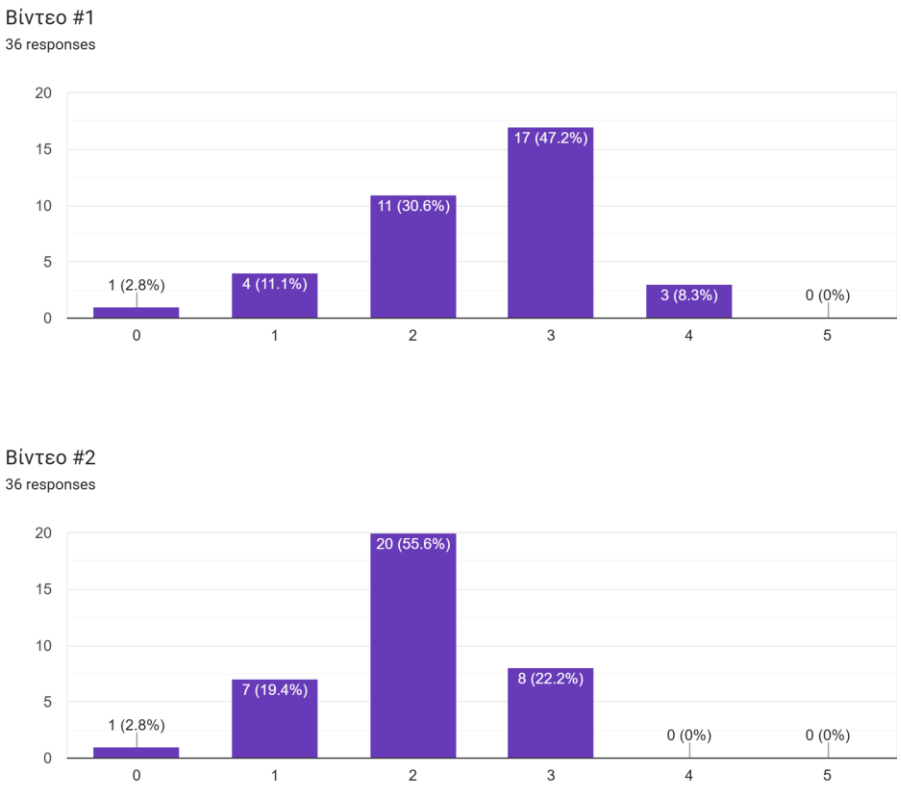
Έπειτα, παίρνοντας από μια στρογγυλοποίηση τις τιμές εξόδου του μοντέλου καταλήγουμε σε αυτό το διάγραμμα όπου είναι πιο διακριτή η σύγκριση με τις επισημειώσεις.



Σχήμα 4: Στρογγυλοποιημένες τιμές πρόβλεψης μοντέλου και σύγκριση με επισημειώσεις

Στην συνέχεια πραγματοποιήθηκε μια έρευνα με σκοπό να ελεγχθεί κατά πόσον το ανθρώπινο αυτί μπορεί να εντοπίσει τις επικαλύψεις των ηχητικών συμβάντων σε ένα ηχητικό απόσπασμα. Συγκεκριμένα, στα παραδείγματα που παρουσιάστηκαν παραπάνω, 36 άτομα κλήθηκαν να καταμετρήσουν τους ταυτόχρονα επικαλυπτόμενους ήχους σε συγκεκριμένο σημείο του ηχητικού αποσπάσματος. Και

στις 2 περιπτώσεις η σωστή απάντηση για το πλήθος των ταυτόχρονων ήχων είναι 3. Στην πρώτη περίπτωση, σχεδόν οι μισοί συμμετέχοντες βρήκαν την σωστή απάντηση με ποσοστό 47% ενώ στην δεύτερη περίπτωση οι συμμετέχοντες απάντησαν σε ποσοστό 88% λανθασμένα. Οι ακριβείς απαντήσεις φαίνονται στο σχήμα παρακάτω. Από τα αποτελέσματα συμπεραίνεται η δυσκολία αναγνώρισης και καταμέτρησης των επικαλυπτόμενων ήχων, καθιστώντας αυτό το έργο απαιτητικό για το ανθρώπινο αυτί.



Σχήμα 5: Απαντήσεις ερωτηματολογίου

3. Σύνοψη και μελλοντικές προεκτάσεις

Συνοψίζοντας, τα μέχρι τώρα αποτελέσματα του μοντέλου φαίνεται να προβλέπουν σε ενθαρρυντικό βαθμό τον σωστό αριθμό επικαλύψεων. Ωστόσο, η έως τώρα απόδοση του μοντέλου περιορίζεται από μερικούς παράγοντες. Αρχικά, το ίδιο το προεκπαιδευμένο μοντέλο είναι σχετικά μικρό, καθώς έπρεπε να μπορεί να

αξιοποιηθεί με τον εξοπλισμό που υπάρχει διαθέσιμος. Επίσης μικρό είναι και το σύνολο δεδομένων, καθώς έπρεπε να ελεγχθεί ένα προς ένα το κάθε ηχητικό αρχείο, διασφαλίζοντας πως δεν υπάρχουν κενά στην αρχή και στο τέλος τους, οδηγώντας σε περιορισμένο πλήθος πρωτόλειων ήχων. Επιπροσθέτως, το μοντέλο εκπαιδεύτηκε σε συνολικά 10,000 (δέκα χιλιάδες) μίξεις ήχων λόγω περιορισμένου εξοπλισμού. Τέλος, λόγω πάλι του εξοπλισμού έπρεπε η εκπαίδευση να περιοριστεί σε πυκνότητα επικάλυψης ήχων με μέγιστο το 5 κάτι που περιορίζει και τις μετέπειτα προβλέψεις σε μέγιστο 5.

Διορθώνοντας τα προβλήματα που αναφέρθηκαν, θα μπορούσε να οδηγήσει σε πολύ καλύτερα αποτελέσματα στις προβλέψεις του μοντέλου.

Μελλοντικά, εκτός από τις αλλαγές που προτάθηκαν, θα άξιζε να προστεθεί και άλλο μοντέλο μηχανικής μάθησης το οποίο θα εκτελεί πολυκάναλη κατηγοριοποίηση των ήχων ή ακόμα και διαχωρισμό των καναλιών.

4. Αναφορές

[1] Mesaros, A., Heittola, T., Virtanen, T., & Plumbley, M. D. (2021). Sound Event Detection: A tutorial. IEEE Signal Processing Magazine, 38, 67-83.

[2] Nolasco, I., Singh, S., Morfi, V., Lostanlen, V., Strandburg-Peshkin, A., Vidaña-Vila, E., Gill, L., Pamuła, H., Whitehead, H., Kiskin, I., Jensen, F. H., Morford, J., Emmerson, M. G., Versace, E., Grout, E., Liu, H., & Ghani, B. (2023). Learning to detect an animal sound from five examples. Ecological Informatics, 102258. doi: 10.1016/j.ecoinf.2023.102258.

[3]Rougui, J. E., Istrate, D., & Souidene, W. (2009). Audio sound event identification for distress situations and context awareness. In 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 3501-3504). Minneapolis, MN, USA. doi: 10.1109/IEMBS.2009.5334581.

[4] Elizalde, B., Deshmukh, S., Al Ismail, M., & Wang, H. (2023, June). Clap learning audio concepts from natural language supervision. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[5] Kong, Q., Chen, K., Liu, H., Du, X., Berg-Kirkpatrick, T., Dubnov, S., & Plumbley, M. D. (2023). Universal source separation with weakly labelled data. arXiv preprint arXiv:2305.07447.

[6] https://freesound.org/ (Οκτώβριος 2024)

[7] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems (NeurIPS).

[8] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.