



Σχολή Θετικών Επιστημών και Τεχνολογίας
Μεταπτυχιακή Εξειδίκευση στα Πληροφοριακά Συστήματα

Διπλωματική Εργασία

Ανάλυση και Ανάπτυξη Τεχνικών Μηχανικής Μάθησης
για τη Δημιουργία Μουσικής από Κείμενο

Βάϊος Παππάς

Επιβλέπων καθηγητής Α: Δημήτριος Καστανιώτης

Επιβλέπων καθηγητής Β: Γεώργιος Ρηγόπουλος

Πάτρα, Ιούνιος 2026

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του/της φοιτητή/φοιτήτριας («συγγραφέας/δημιουργός») που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης ο συγγραφέας/δημιουργός εκχωρεί στο ΕΑΠ, μη αποκλειστική άδεια χρήσης του δικαιώματος αναπαραγωγής, προσαρμογής, δημόσιου δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσής τους διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος και για όλο το χρόνο διάρκειας των δικαιωμάτων πνευματικής ιδιοκτησίας. Η ανοικτή πρόσβαση στο πλήρες κείμενο για μελέτη και ανάγνωση δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, αποθήκευση, πώληση, εμπορική χρήση, μετάδοση, διανομή, έκδοση, εκτέλεση, «μεταφόρτωση» (downloading), «ανάρτηση» (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού. Ο συγγραφέας/δημιουργός διατηρεί το σύνολο των ηθικών και περιουσιακών του δικαιωμάτων.



Ανάλυση και Ανάπτυξη Τεχνικών Μηχανικής Μάθησης
για τη Δημιουργία Μουσικής από Κείμενο

Βάιος Παπάς

Επιτροπή Επίβλεψης Διπλωματικής Εργασίας

Επιβλέπων Καθηγητής:

Συν-Επιβλέπων Καθηγητής:

Δημήτριος Καστανιώτης

Γεώργιος Ρηγόπουλος

Πάτρα, Ιούνιος 2026

«Ευχαριστώ την οικογένεια μου και τη σύντροφο μου που με έχουν στηρίζει στην ακαδημαϊκή μου πορεία στο ΕΑΠ καθώς και το ίδιο το ίδρυμα που μου έδωσε την ευκαιρία να παρακολουθήσω το αγαπημένο μου μεταπτυχιακό πρόγραμμα.»

Περίληψη

Αυτή η εργασία παρουσιάζει μια συστηματική και συγκριτική ανάλυση των αρχιτεκτονικών βαθιάς μάθησης και των αλγοριθμικών διαδικασιών που απαιτούνται για τη μετατροπή μη δομημένων προτροπών φυσικής γλώσσας σε συνεχείς ακουστικές κυματομορφές. Αναλύονται πέντε αλγοριθμικά στάδια: οι μηχανισμοί διακριτοποίησης και ενσωμάτωσης κειμένου, η συμπίεση και διακριτοποίηση νευρωνικού ήχου μέσω κβαντοποίησης διανυσμάτων, τα διασταυρούμενα επίπεδα δομικής χαρτογράφησης όπως μηχανισμοί βελτιστοποίησης που χρησιμοποιούν αλγόριθμους βασισμένους σε διαβάθμιση και οι χρονοπρογραμματιστές δειγματοληψίας συμπερασμάτων χρόνου εκτέλεσης. Επιπλέον, εξετάζονται τα δύο κυρίαρχα γενετικά παραδείγματα: η Αυτοπαλίνδρομη Μοντελοποίηση Γλώσσας και τα Μοντέλα Λανθάνουσας Διάχυσης. Αναλύεται πώς τα αυτοπαλίνδρομα συστήματα αντιμετωπίζουν τη δημιουργία μουσικής ως ένα διαδοχικό πρόβλημα πρόβλεψης διακριτικών σε σχέση με αρχιτεκτονικές που βασίζονται στη διάχυση και συνθέτουν ήχο αφαιρώντας επαναληπτικά τον θόρυβο από έναν λανθάνοντα χώρο μέσω ντετερμινιστικών χρονοπρογραμματιστών. Κάθε παράδειγμα αξιολογείται χρησιμοποιώντας αντικειμενικές μετρήσεις αναφοράς όσον αφορά την ακουστική πιστότητα ενός παραγόμενου ήχου, βαθμολογίες εξάρτησης και σύνδεσης του ήχου με την φυσική γλώσσα και βαθμολογίες μέσης γνώμης που αξιολογούν την αντιληπτική ποιότητα ενός μηχανικά δημιουργημένου ήχου.

Λέξεις – Κλειδιά

Δημιουργία Κειμένου σε Ήχο, Αρχιτεκτονική Μετασχηματιστών, Μοντέλα Λανθάνουσας Διάχυσης, Μηχανισμός Αυτοπροσοχής, Κβαντοποίηση Διανυσμάτων, Νευρωνικοί Κωδικοποιητές Ήχου, Πολυτροπική Μηχανική Μάθηση

Analysis and Development of Machine Learning Algorithms for Text-to-Music Generation

Vaios Pappas

Abstract

This paper presents a systematic and comparative analysis of deep learning architectures and algorithmic processes required to transform unstructured natural language prompts into continuous acoustic waveforms. Five algorithmic stages are analyzed: tokenization and text embedding mechanisms, neural audio compression and discretization via vector quantization, cross-layer structural mapping such as optimization mechanisms using gradient-based algorithms, and runtime inference sampling schedulers. In addition, the two dominant genetic paradigms are examined: Autoregressive Language Modeling and Latent Diffusion Models. It analyzes how autoregressive systems approach music creation as a sequential discriminator prediction problem relative to diffusion-based architectures that synthesize sound by iteratively removing noise from a latent space via deterministic schedulers. Each example is evaluated using objective benchmark measures of the acoustic fidelity of a produced sound, scores of dependence and connection of the sound to natural language, and mean opinion scores that assess the perceptual quality of a mechanically generated sound.

Keywords

Text-to-Audio Generation, Transformer Architecture, Latent Diffusion Models, Self-Attention Mechanism, Vector Quantization, Neural Audio Codecs, Multi-Modal Machine Learning

Κατάλογος περιεχομένων

Περίληψη.....	v
Abstract.....	vi
Συντομογραφίες & Ακρωνύμια	xiii
1. Εισαγωγή.....	1
1.1 Η Εξέλιξη της Γενετικής Τεχνητής Νοημοσύνης.....	1
1.2 Τα Υπολογιστικά και Δομικά Εμπόδια της Σύνθεσης Κειμένου σε Ήχο	3
1.2.1 Υπολογιστικά Εμπόδια.....	4
1.2.2 Δομικά Εμπόδια: Ασυμμετρία Πληροφοριών και Αντιστοίχιση Ένα-προς-Πολλά	5
1.3 Σκοπός της εργασίας.....	6
1.4 Δομή της Εργασίας	8
2. Θεωρητικό Υπόβαθρο & Βασικά Στοιχεία.....	10
2.1 Επεξεργασία Φυσικής Γλώσσας & Ενσωμάτωση Κειμένου	10
2.1.1 Μηχανική Τεμαχιοποίησης	10
2.1.2 Στατικοί Χώροι Ενσωμάτωσης	11
2.1.3 Χώροι Ενσωμάτωσης με βάση τα Συμφραζόμενα	12
2.2 Ψηφιακή Επεξεργασία Σήματος (DSP) & Αναπαραστάσεις Ήχου	14
2.2.1 Το Πεδίο του Χρόνου έναντι του Πεδίου της Συχνότητας	14
2.2.2 Ο Μετασχηματισμός Fourier Βραχύχρονου (STFT)	16
2.2.3 Φασματογράμματα Mel και Λογαριθμική Συμπίεση	19
2.3 Αρχιτεκτονικές Βαθιάς Μάθησης για Χαρτογράφηση Ακολουθιών.....	23
2.3.1 Επαναληπτικές Προσεγγίσεις: RNN, LSTM και το Sequential Bottleneck.....	23
2.3.2 Η Αρχιτεκτονική του Transformer και η Μηχανική της Αυτοπροσοχής	28
2.3.3 Τοπολογίες Προσοχής: Αιτιώδης Διαμόρφωση Μάσκας έναντι Αμφίδρομου Πλαισίου	31
2.3.4 Ακουστική Προσαρμογή: Χαρτογράφηση Ακολουθίας με Επαυξημένη Συνέλιξη	32
3. Ο αλγοριθμικός αγωγός για τη μετατροπή κειμένου σε ήχο	36
3.1 Εξαγωγή Χαρακτηριστικών Front-End (Αλγόριθμοι Κειμενοποίησης).....	38
3.1.1 Αλγοριθμική Τεμαχιοποίηση και Χαρτογράφηση Λεξιλογίου	39
3.1.2 Αρχιτεκτονικά Παραδείγματα Κωδικοποίησης Κειμένου.....	40

3.1.3 Η στρατηγική των παγωμένων βαρών και της εξαγωγής χαρακτηριστικών	43
3.1.4 Αγωγοί Επαύξησης Κειμένου και Παραγωγής Λεζαντών (Captioning)	44
3.1.5 Μηχανική Άνευ Όρων Προετοιμασίας και Καθοδήγηση Χωρίς Ταξινομητές (CFG)	45
3.2 Διακριτοποίηση και Ακουστική Συμπύεση (Νευρωνικοί Κωδικοποιητές Ήχου)	46
3.2.1 Η θεμελιώδης αρχιτεκτονική των νευρωνικών κωδικοποιητών ήχου	46
3.2.2 Μηχανική Κβαντοποίησης Διανυσμάτων (VQ).....	50
3.2.3 Το Τείχος Κλιμάκωσης του Ρυθμού Μετάδοσης και η Κβαντοποίηση Υπολειμματικών Διανυσμάτων (RVQ)	52
3.2.4 Πεπερασμένη Βαθμωτή Κβαντοποίηση (FSQ)	56
3.2.5 Μετρήσεις Σταθεροποίησης: Κανονικοποίηση Διανύσματος Υπερσφαίρας	57
3.2.6 Βελτιστοποίηση Βιβλίου Κωδικών	59
3.3 Διατροφική Δομική Χαρτογράφηση (Transformer/Attention Core).....	60
3.3.1 Μηχανική Συνθήκης Διασταυρούμενης Προσοχής	61
3.3.2 Διατροφική Χαρτογράφηση σε Μοντέλα Λανθάνουσας Διάχυσης (Latent Diffusion)	64
3.3.3 Διαμορφώσεις Καθολικής Προετοιμασίας: Προσαρμοστική Κανονικοποίηση Επιπέδου (adaLN-Modulation)	66
3.3.4 Τυποποίηση Μήκους Ακολουθίας: Αναδειγματοληψία Αντιλήπτη (Perceiver Resampling)	69
3.3.5 Διατροφική Χαρτογράφηση σε Αυτοπαλίνδρομα Μοντέλα (Διαπλοκή Ακολουθιών)	71
3.4 Βελτιστοποίηση Συστήματος (Η Μηχανή Εκπαίδευσης).....	72
3.4.1 Η Μηχανή Προσαρμοστικής Κλίσης: Adam and AdamW.....	73
3.4.2 Στόχοι Κατηγορικής και Συνεχούς Ζημίας.....	75
3.4.3 Μοντέλο ταύτισης ροής (Flow Matching).....	78
3.4.4 Αρχιτεκτονική Απώλειας Αντιστοίχισης Χαρακτηριστικών Πολλαπλών Κλιμάκων	79
3.5 Σύνθεση Ήχου Back-End: Ανακατασκευή της Ακουστικής Κυματομορφής κατά τον Χρόνο Εκτέλεσης.....	83
3.5.1 Χρονοπρογραμματιστές Δειγματοληψίας και Βρόχοι Παρέκτασης (Sampling Schedulers)	84
3.5.2 Νευρωνικοί Ακουστικοί Αποκωδικοποιητές (Neural Vocoder)	85

3.5.3 Στοχαστική Δειγματοληψία και Κλάδεμα Token	86
3.5.4 Ο Βρόχος Συμπερασμάτων Διπλής Διέλευσης και Καθοδήγηση Χωρίς Ταξινομητή (CFG)	88
3.5.5 Χρονοπρογραμματιστές για Προφίλ Συνεχούς Διάχυσης: Ντετερμινιστική Αποθορυβοποίηση	90
3.5.6 Επιτάχυνση Τροχιάς Υψηλής Τάξης: DPM-Solver++	92
3.5.7 Η Τελική Αποσυμπίεση: Σύνθεση Κυματομορφής μέσω Νευρωνικών Vocoderς	94
4. Συγκριτική Ανάλυση Γενετικών Παραδειγμάτων.....	98
4.1 Αυτοπαλίνδρομη Μοντελοποίηση Γλώσσας και Διαδοχική Παραγωγή Ακουστικών Διακριτικών	98
4.1.1 Μαθηματικό Πλαίσιο και Εξάρτηση Ακολουθίας	98
4.1.2 Αναλυτική αξιολόγηση: Πλεονεκτήματα και Μειονεκτήματα.....	100
4.2 Μοντέλα Λανθάνουσας Διάχυσης: Συνεχής Ακουστική Αποθορυβοποίησης σε Λανθάνοντες Χώρους.....	103
4.2.1 Μαθηματικό Πλαίσιο και η Τροχιά Αποθορυβοποίησης	104
4.2.2 Αναλυτική Αξιολόγηση: Πλεονεκτήματα και Μειονεκτήματα	106
5. Εμπειρικά Σημεία Αναφοράς & Μεθοδολογία Αξιολόγησης.....	110
5.1 Αντικειμενικές Μετρήσεις Αξιολόγησης	110
5.1.1 Απόσταση Ήχου Fréchet (Fréchet Audio Distance - FAD)	110
5.1.2 Βαθμολογία CLAP (Contrastive Language-Audio Pretraining Score)	112
5.2 Υποκειμενικές Μετρήσεις Αξιολόγησης: Μέση Βαθμολογία Γνώμης (MOS)	114
5.2.1 Μηχανική και Διαδικασία των Δοκιμών MOS	114
5.2.2 Υποκειμενική Αξιολόγηση Διπλού Άξονα	116
5.3 Υποκειμενικές Μετρήσεις Πολλαπλών Συνθηκών MUSHRA	118
5.4 Εκπαίδευση μοντέλων και σύγκριση αποτελεσμάτων(FAD-CLAP).....	119
5.4.1 Εκπαίδευση με βάση την αρχιτεκτονική Transformer.....	120
5.4.2 Εκπαίδευση με βάση την αρχιτεκτονική Diffusion.....	120
5.4.3 Σύγκριση μετρικών για κάθε αρχιτεκτονική.....	121
6. Συμπεράσματα,περιορισμοί και μελλοντικοί στόχοι.....	124
6.1 Συμπεράσματα.....	124
6.2 Περιορισμοί.....	125
6.2.1 Το πρόβλημα της συνοχής σε μεγάλο εύρος: Γιατί τα μοντέλα δυσκολεύονται να διατηρήσουν ένα συνεπές μουσικό θέμα για αρκετά λεπτά.....	125

<u>6.2.2 Περιορισμοί Υπολογισμού και Υλικού.....</u>	<u>126</u>
<u>6.3 Προοπτικές για τις Μελλοντικές Κατευθύνσεις της Έρευνας</u>	<u>128</u>
<u>Βιβλιογραφία.....</u>	<u>131</u>
<u>Παράρτημα Α: Κώδικες Υλοποίησης Εκπαίδευσης Μοντέλων.....</u>	<u>140</u>

Κατάλογος Εικόνων/Σχημάτων

Εικόνα 1: Μετατροπή Ήχου από Κείμενο.....	3
Εικόνα 2: Time Domain to Frequency Domain.....	16
Εικόνα 3: Διάγραμμα Μηχανισμού Ολίσθησης Παραθύρου	17
Εικόνα 4: Mel filterbank triangular filters diagram.....	20
Εικόνα 5: Log-Mel Spectrogram.....	22
Εικόνα 6 :Πειραματικά αποτελέσματα σύγκλισης της εκπαίδευσης σε απλά RNN	25
Εικόνα 7: Transformer-Model Architecture.....	29
Εικόνα 8: Conformer encoder model architecture.....	34
Εικόνα 9: Σχηματική περιγραφή του αλγοριθμικού αγωγού απο κείμενο σε ήχο.....	38
Εικόνα 10: Μετατροπή προτροπής σε tokens.....	39
Εικόνα 11: Αρχιτεκτονική και διαδικασία λειτουργίας του μοντέλου CLAP	41
Εικόνα 12: Αρχιτεκτονική Κωδικοποιητή-Αποκωδικοποιητή.....	47
Εικόνα 13: Αρχιτεκτονική του νευρωνικού μοντέλου SoundStream	49
Εικόνα 14: Σχήμα περιγραφής του VQ-VAE.....	51
Εικόνα 15 : Αρχιτεκτονική του νευρωνικού κωδικοποιητή ENCODEC	55
Εικόνα 16: Scaled Dot-Product Attention.....	62
Εικόνα 17: Φάσεις Αποθρονοποίησης.....	65
Εικόνα 18: Η αρχιτεκτονική Diffusion Transformer (DiT).....	68
Εικόνα 19: Αρχιτεκτονική ροή του Perceiver με τη χρήση επαναληπτικής προσοχής	70
Εικόνα 20 : Αρχιτεκτονική των δύο βασικών υποδικτύων διακριτών (sub-discriminators) του πλαισίου HiFi-GAN	80
Εικόνα 21: Αρχιτεκτονική της γεννήτριας (generator) του HiFi-GAN	96

Κατάλογος Πινάκων

Πίνακας 1: Πίνακας Βαθμολογίας Likert-Αξιολόγηση MOS.....	115
Πίνακας 2: Πίνακας αξιολόγησης μετρικών FAD-CLAP.....	122

Συντομογραφίες & Ακρωνύμια

1. Εισαγωγή

1.1 Η Εξέλιξη της Γενετικής Τεχνητής Νοημοσύνης

Η γενετική τεχνητή νοημοσύνη έχει αναπτυχθεί ριζικά την τελευταία δεκαετία με αφετηρία την Επεξεργασία Φυσικής Γλώσσας (NLP). Εκεί, η αρχιτεκτονική Transformer και οι νόμοι νευρωνικής κλιμάκωσης (scaling laws) άλλαξαν ριζικά το τοπίο, επιτρέποντας στις μηχανές να αναπαράγουν τον ανθρώπινο λόγο με εξαιρετική ευχέρεια. Αρχικά, η μοντελοποίηση της γλώσσας παρέμενε μονοτροπική. Οι αλγόριθμοι εστίαζαν σε αυτοπαλίνδρομους στόχους ή στην πρόβλεψη καλυμμένων λέξεων (masked language modeling), μετατρέποντας αυτούσια μια ακολουθία κειμένου σε μια άλλη. Η επιτυχία αυτών των μεγάλων γλωσσικών μοντέλων ανέδειξε μια νέα πραγματικότητα: το κείμενο λειτουργεί πλέον ως μια καθολική, εξαιρετικά πυκνή σημασιολογική διεπαφή, ικανή να κωδικοποιεί σύνθετες ανθρώπινες προθέσεις.

Η διατροπική (cross-modal) χωρική παραγωγή αποτέλεσε το επόμενο λογικό βήμα της έρευνας, με αιχμή του δόρατος τη σύνθεση εικόνας από κείμενο. Μέσα από πλαίσια αντιπαραθετικής μάθησης (contrastive learning) όπως το CLIP, οι ερευνητές ευθυγράμμισαν τους γλωσσικούς και οπτικούς χώρους χαρακτηριστικών. Αυτή η ευθυγράμμιση επέτρεψε την άμεση προβολή των σημασιολογικών αναπαραστάσεων του κειμένου στους γενετικούς αγωγούς. Στη συνέχεια, αρχιτεκτονικές όπως οι Μεταβλητοί Αυτόματοι Κωδικοποιητές (VAEs), τα Γενετικά Αντιπαραθετικά Δίκτυα (GANs) και τα Λανθάνοντα Μοντέλα Διάχυσης (LDMs) μετέτρεψαν αυτούς τους κοινούς χώρους σε λεπτομερείς συστοιχίες pixel. Η συγκεκριμένη εξέλιξη απέδειξε στην πράξη ότι μια αφηρημένη κειμενική προτροπή (prompt) αρκεί για να καθοδηγήσει τη σύνθεση νέων, δομικά συνεκτικών οπτικών δεδομένων. Η σύνθεση ήχου και μουσικής υψηλής πιστότητας παραμένει το πιο απαιτητικό πεδίο για την πολυτροπική μηχανική μάθηση. Τα υπολογιστικά και δομικά εμπόδια εδώ διαφέρουν ριζικά από εκείνα του κειμένου ή των εικόνων:

- **Ακραία Χρονική Πυκνότητα:** Ένα κείμενο περιλαμβάνει μερικές εκατοντάδες tokens και μια εικόνα μερικές χιλιάδες pixels. Αντίθετα, ο ακατέργαστος ψηφιακός Διπλωματική Εργασία

ήχος σε ποιότητα CD απαιτεί 44.1 kHz —δηλαδή 44.100 σημεία δεδομένων για κάθε δευτερόλεπτο παραγωγής. Ένα τυπικό τρίλεπτο μουσικό κομμάτι ξεπερνά τα 7,9 εκατομμύρια διαδοχικά δείγματα, προκαλώντας τεράστια υπολογιστική ασφυξία στα παραδοσιακά μοντέλα ακολουθιών.

- **Ιεραρχική Δομή Πολλαπλών Κλιμάκων:** Η μουσική κρύβει μια εξαιρετικά πολύπλοκη, πολυεπίπεδη ιεραρχία. Στη μικροκλίμακα (χιλιοστά του δευτερολέπτου), το μοντέλο διαμορφώνει κυματομορφές, ηχοχρώματα και υφές οργάνων. Στη μεσοκλίμακα (δευτερόλεπτα), οργανώνει νότες, ρυθμούς και τοπικές μελωδίες. Στη μακροκλίμακα (λεπτά), οφείλει να διατηρεί τη δομική συνοχή, το μουσικό είδος και τη θεματική εξέλιξη του κομματιού.
- **Ακουστική Υπέρθεση:** Στις εικόνες, τα αντικείμενα καταλαμβάνουν σαφή, διακριτά χωρικά όρια. Στη μουσική, οι συχνότητες συμπλέκονται σε μια σύνθετη υπέρθεση. Ένα μοναδικό ηχητικό δείγμα περιλαμβάνει ταυτόχρονα τις μείξεις των φωνητικών, των κρουστών, του μπάσου και της αρμονίας. Το μοντέλο καλείται να αποτυπώσει αυτές τις πυκνές, πολυοργανικές αλληλεπιδράσεις μέσα από μια μονοδιάστατη ροή συνεχών μεταβλητών.

Η ανάγκη για μια κριτική ανάλυση αυτών ακριβώς των προκλήσεων αποτελεί το βασικό κίνητρο της παρούσας εργασίας. Η σύνθεση μουσικής από κείμενο ξεπερνά τα όρια ενός απλού μηχανικού επιτεύγματος· δοκιμάζει σε βάθος την ικανότητα μιας αρχιτεκτονικής να μεταφράζει την αφηρημένη ανθρώπινη σημασιολογία σε ακριβείς ακουστικές δομές, όπως για παράδειγμα:

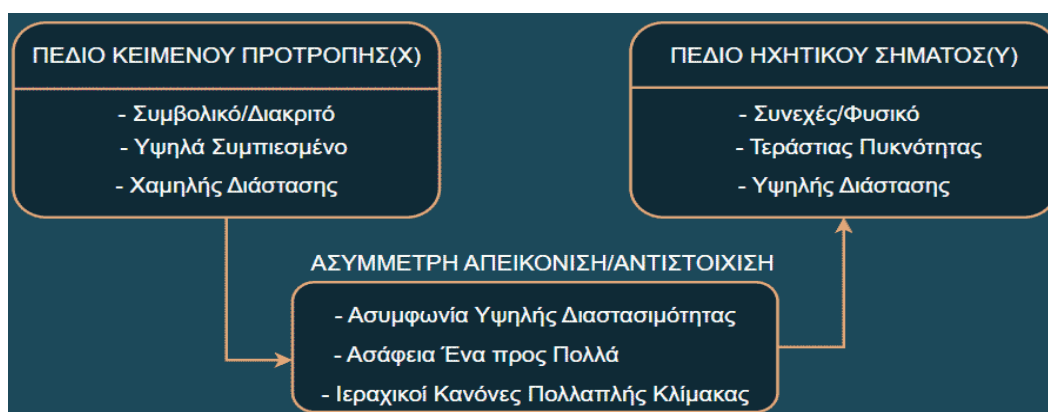
"Ένα μελαγχολικό riff ακουστικής κιθάρας σε Μι ελάσσονα με έναν αργό, δυναμικό ρυθμό τυμπάνου lo-fi."

Τα δυναμικά μοντέλα βαθιάς μάθησης αντικαθιστούν πλέον την παραδοσιακή, χειροκίνητη σύνθεση, αναδιαμορφώνοντας ριζικά την παραγωγή μουσικής, την επεξεργασία σήματος και τα διαδραστικά μέσα.

Στο επίκεντρο αυτής της εργασίας βρίσκεται η συγκριτική αξιολόγηση των αλγοριθμικών αγωγών, των μεθόδων συμπίεσης και των βασικών αρχιτεκτονικών, συγκρίνοντας την αυτοπαλίνδρομη μοντελοποίηση (autoregressive modeling) με τη λανθάνουσα διάχυση (latent diffusion). Η προσέγγιση αυτή χαρτογραφεί τις αναγκαίες ισορροπίες ανάμεσα στην υπολογιστική ισχύ, την πιστότητα του ήχου και τη μακροπρόθεσμη δομική συνοχή. Στόχος είναι να προσφέρει μια σαφή μαθηματική και δομική ταξινόμηση των αλγορίθμων που μετατρέπουν το γραπτό λόγο σε ζωντανά ηχοτοπία.

1.2 Τα Υπολογιστικά και Δομικά Εμπόδια της Σύνθεσης Κειμένου σε Ήχο

Η σύνθεση κειμένου σε ήχο στοχεύει στη δημιουργία μιας αυτοματοποιημένης, παραγωγικής χαρτογράφησης. Μεταφράζει ένα διακριτό πεδίο πηγής φυσικής γλώσσας X σε ένα συνεχές πεδίο στόχο ακουστικών σημάτων υψηλής πιστότητας Y , όπως παρουσιάζεται στην [Εικόνα 1](#). Η διατροπική παραγωγή λειτουργεί ήδη σταθερά σε χωρικούς τομείς, όπως η σύνθεση κειμένου σε εικόνα. Αντίθετα, η μετατροπή κειμενικών περιγραφών σε συνεκτικές ηχητικές κυματομορφές εισάγει ασύμμετρους περιορισμούς.



Εικόνα 1: Μετατροπή Ήχου από Κείμενο

1.2.1 Υπολογιστικά Εμπόδια

Η βαθιά αριθμητική αναντιστοιχία στην πυκνότητα δεδομένων και το μήκος ακολουθίας μεταξύ των δύο τρόπων (modalities) αποτελεί το κύριο υπολογιστικό εμπόδιο στη μοντελοποίηση.

- **Αναντιστοιχία μήκους ακολουθίας:** Μερικές δεκάδες λέξεις στην επεξεργασία φυσικής γλώσσας μεταφράζονται σε μόλις 10^1 έως 10^2 διακριτικά υπολέξεων (tokens). Ο ψηφιακός ήχος, όμως, αποτελεί μια συνεχή φυσική κυματομορφή. Η επαγγελματική ακουστική πιστότητα απαιτεί συχνότητα δειγματοληψίας τουλάχιστον 44,1 kHz. Έτσι, η παραγωγή ενός και μόνο λεπτού στερεοφωνικού ήχου απαιτεί την επεξεργασία άνω των $2,64 \cdot 10^6$ συνεχών μεταβλητών ανά κανάλι. Η αριθμητική αυτή απόσταση είναι χαοτική.
- **Το σημείο συμφόρησης στην πολυπλοκότητα προσοχής:** Ένα άμεσο μαθηματικό σημείο αποτυχίας προκύπτει από την απευθείας εισαγωγή αυτών των ακολουθιών σε σύγχρονα μοντέλα. Ο τυπικός μηχανισμός αυτοπροσοχής (self-attention) των αρχιτεκτονικών Transformer κλιμακώνεται τετραγωνικά κατά $O(N^2)$, ως προς το μήκος ακολουθίας N . Ο υπολογισμός της συνολικής αυτοπροσοχής σε εκατομμύρια ακατέργαστα δείγματα ήχου ξεπερνά τη φυσική χωρητικότητα μνήμης των σύγχρονων GPU. Το γεγονός αυτό επιβάλλει τη χρήση επιθετικών πλαισίων συμπίεσης με απώλειες.
- **Υπερφόρτωση πληροφοριών και εξαγωγή αναπαράστασης:** Ένα μονοδιάστατο κύμα πίεσης εγκλωβίζει θόρυβο περιβάλλοντος, αντανakλάσεις δωματίου, παρεμβολές οργάνων και διακυμάνσεις φάσης. Τα επίπεδα του κωδικοποιητή αναλώνουν τεράστια υπολογιστική ισχύ για να φιλτράρουν αυτά τα περιττά ακουστικά στοιχεία. Στόχος του μοντέλου είναι να απομονώσει τη δομή χωρίς να θυσιάσει τις λεπτές ακουστικές υφές που καθορίζουν την τελική ποιότητα.

1.2.2 Δομικά Εμπόδια: Ασυμμετρία Πληροφοριών και Αντιστοίχιση Ένα-προς-Πολλά

Βαθιές αρχιτεκτονικές αντιφάσεις στον τρόπο κωδικοποίησης κειμένου και ήχου δημιουργούν πρόσθετα προβλήματα, πέρα από τους περιορισμούς του υλικού:

- **Το μη-ντετερμινιστικό πρόβλημα «ένα-προς-πολλά»:** Το κείμενο αποτελεί ένα εξαιρετικά συμβολικό, αφηρημένο και συμπίεσμένο μέσο. Μια προτροπή (prompt) όπως «Μια αργή ακουστική μελωδία κιθάρας» αφήνει άπειρες ακουστικές μεταβλητές απροσδιόριστες. Η εξέλιξη των συγχορδιών, το τέμπο, το υλικό του σώματος της κιθάρας, η δύναμη των χορδών και η χωρική μίξη παραμένουν άγνωστα στοιχεία. Συνεπώς, ένα μοναδικό σημείο στον σημασιολογικό χώρο του κειμένου αντιστοιχεί σε μια άπειρη πολλαπλότητα έγκυρων ακουστικών παραλλαγών. Το μοντέλο οφείλει να χαρτογραφήσει μια σύνθετη κατανομή πιθανοτήτων για να επιλέξει μια εύλογη ακουστική πραγματικότητα.
- **Αφηρημένη προς φυσική διασταυρούμενη ευθυγράμμιση:** Σημασιολογική σημασία υψηλού επιπέδου χαρακτηρίζει τα διακριτικά κειμένου, την ίδια στιγμή που οι ακατέργαστες τιμές ήχου ορίζονται από φυσικά χαρακτηριστικά χαμηλού επιπέδου (πλάτος και συχνότητα). Η απευθείας αντιστοίχιση ενός αφηρημένου συναισθηματικού επιθέτου όπως «ζοφερό» ή μιας πολιτισμικής ετικέτας όπως «lo-fi» σε μια ψηφιακή κυματομορφή είναι αδύνατη. Απαιτείται η εκμάθηση ενός ενδιάμεσου λανθάνοντος χώρου (latent space). Εκεί, η γλωσσική αισθητική μεταφράζεται τελικά σε φυσικές παραμέτρους ήχου, όπως η εντοπισμένη εξασθένηση υψηλών συχνοτήτων ή συγκεκριμένα αρμονικά διαστήματα ελάσσονος κλίμακας.
- **Συνοχή φάσεων και συνέργεια μακράς εμβέλειας:** Τα δείγματα ήχου βασίζονται σε ακριβείς ευθυγραμμίσεις φάσεων στον χρόνο για να παράγουν καθαρούς τόνους. Αυτό έρχεται σε πλήρη αντίθεση με τα pixel, όπου οι γειτονικές τιμές απλώς υπαγορεύουν στατικά χρώματα. Μια απειροελάχιστη μετατόπιση ελάχιστων δειγμάτων αρκεί για να μετατρέψει μια άψογη μουσική συγχορδία σε καταστροφική ακύρωση φάσης ή σκληρή παραμόρφωση. Παράλληλα, το σύστημα οφείλει να επιβάλλει άκαμπτους μακροδομικούς κανόνες, διατηρώντας για παράδειγμα έναν

ρυθμό τυμπάνων αυστηρά κλειδωμένο σε ένα πλέγμα 120 BPM κατά τη διάρκεια εκατομμυρίων δειγμάτων.

1.3 Σκοπός της εργασίας

Κεντρική επιδίωξη αυτής της εργασίας είναι η διεξαγωγή μιας ολοκληρωμένης, μαθηματικά αυστηρής και δομικής αξιολόγησης των σύγχρονων τεχνικών μηχανικής μάθησης που εφαρμόζονται στην διατροφική παραγωγή κειμένου σε μουσική (text-to-music). Η έρευνα αυτή απορρίπτει την αντιμετώπιση των γενετικών μοντέλων ως μονολιθικών οντοτήτων. Αντίθετα, αποδομεί τα σύγχρονα συστήματα με σκοπό να απομονώσει, να συγκρίνει και να αναλύσει τα μεμονωμένα αλγοριθμικά επίπεδα που καθιστούν δυνατή τη διατροφική σύνθεση.

Για την υλοποίηση αυτού του σκοπού, η ερευνητική διαδικασία εξειδικεύεται στους ακόλουθους λειτουργικούς στόχους:

- **Αποδόμηση του αγωγού παραγωγής από άκρο σε άκρο (End-to-End Pipeline):** Χαρτογράφηση και ανάλυση της πλήρους λειτουργικής ακολουθίας των συστημάτων κειμένου σε ήχο. Το ενδιαφέρον εστιάζεται στην αξιολόγηση του τρόπου με τον οποίο η διακριτή προεπεξεργασία κειμένου (Byte Pair Encoding) αλληλεπιδρά με τα νευρωνικά επίπεδα συμπίεσης ήχου (Residual Vector Quantization), διαμορφώνοντας μια βιώσιμη υπολογιστική διεπαφή πριν από την έναρξη της παραγωγής.
- **Αξιολόγηση δομικών αρχιτεκτονικών (Επανάληψη έναντι Προσοχής):** Διεξαγωγή αναλυτικής σύγκρισης ανάμεσα στα ιστορικά μοντέλα ακολουθίας (RNN, LSTM) και τις σύγχρονες αρχιτεκτονικές προσοχής (Transformers). Ο στόχος αυτός επικεντρώνεται στο πώς οι μαθηματικοί μηχανισμοί της αυτοπροσοχής πολλαπλών κεφαλών (Multi-Head Self-Attention) επιλύουν τα διαδοχικά σημεία συμφόρησης και το πρόβλημα της εξαφάνισης της διαβάθμισης (vanishing gradients), διασφαλίζοντας τη μουσική συνοχή μεγάλης εμβέλειας.

- **Σύγκριση βασικών γενετικών παραδειγμάτων (Αυτοπαλινδρόμηση έναντι Διάχυσης):** Συστηματική κριτική των δύο κυρίαρχων γενετικών πλαισίων στο τοπίο της σύνθεσης ήχου:
 - Αυτοπαλινδρομική Μοντελοποίηση Γλώσσας(Transformer): Εξέταση των μηχανισμών διαδοχικής πρόβλεψης διακριτών ακουστικών tokens.
 - Μοντέλα Λανθάνουσας Διάχυσης(Diffusion): Διερεύνηση της επαναληπτικής αφαίρεσης θορύβου από μια λανθάνουσα αναπαράσταση με τη χρήση συνεχών χρονοπρογραμματιστών (όπως DDPM και DDIM).
- **Ανάλυση της μηχανικής χαρτογράφησης του λανθάνοντος χώρου (Latent Space):** Διερεύνηση του τρόπου με τον οποίο οι αφηρημένες γλωσσικές εκφράσεις (σημασιολογικά prompts) προβάλλονται και ευθυγραμμίζονται με φυσικές ακουστικές μεταβλητές (συχνότητες, πλάτη και ευθυγραμμίσεις φάσης). Η ανάλυση περιλαμβάνει την εξέταση των μηχανισμών διασταυρούμενης προσοχής (cross-attention) που επιτρέπουν στα διανύσματα κειμένου να κατευθύνουν τη διαδικασία σύνθεσης.
- **Ποσοτικοποίηση της απόδοσης μέσω τυποποιημένων δεικτών:** Μέτρηση των δομικών συμβιβασμών (trade-offs) κάθε τεχνικής μηχανικής μάθησης με τη χρήση μιας δέσμης αντικειμενικών και υποκειμενικών μετρήσεων. Συγκεκριμένα, τα αποτελέσματα των μοντέλων αξιολογούνται με βάση την Απόσταση Ήχου Fréchet (FAD) για την απόλυτη ακουστική πιστότητα, τις βαθμολογίες CLAP για τη σημασιολογική ευθυγράμμιση κειμένου-ήχου, και τη Μέση Βαθμολογία Γνώμης (MOS) για την ποιοτική ανθρώπινη αξιολόγηση.

1.4 Δομή της Εργασίας

Η παρούσα διατριβή διαρθρώνεται σε έξι ξεχωριστά κεφάλαια. Η δομή αυτή καθοδηγεί συστηματικά τον αναγνώστη από τα θεωρητικά θεμέλια έως την ολοκληρωμένη αλγοριθμική ανάλυση, καταλήγοντας στην εμπειρική αξιολόγηση των τεχνικών διατροφικής σύνθεσης, όπως παρουσιάζεται συνοπτικά στα παρακάτω τέσσερα κεφάλαια:

Κεφάλαιο 2: Θεωρητικό Υπόβαθρο & Βασικά Στοιχεία

Το κεφάλαιο αυτό θέτει τις επιστημονικές βάσεις που υποστηρίζουν και τις δύο μεθόδους σύνθεσης.

- **Επεξεργασία Φυσικής Γλώσσας (NLP):** Αναλύονται οι στατικές και οι συμφραζόμενες (contextual) ενσωματώσεις κειμένου (embeddings).
- **Ψηφιακή Επεξεργασία Σήματος (DSP):** Εξετάζονται θεμελιώδεις έννοιες, όπως ο Μετασχηματισμός Fourier Βραχύχρονου (STFT) και τα φασματογραφήματα Mel.
- **Μοντελοποίηση Ακολουθιών:** Παρουσιάζεται η εξέλιξη των αρχιτεκτονικών, αντιπαραβάλλοντας άμεσα τους επαναλαμβανόμενους μηχανισμούς (RNN και LSTM) με το μαθηματικό πλαίσιο της αυτοπροσοχής πολλαπλών κεφαλών (Multi-Head Self-Attention).

Κεφάλαιο 3: Ο Αλγοριθμικός αγωγός για τη μετατροπή κειμένου σε ήχο

Αυτό το κεφάλαιο αποτελεί τη δομική καρδιά της μηχανικής ανάλυσης. Αποδομεί τη διαδικασία παραγωγής ήχου από κείμενο (text-to-audio) σε μια ενοποιημένη γραμμή συναρμολόγησης πολλαπλών σταδίων.

- **Front-end επεξεργασία:** Λεπτομερής ανάλυση των μηχανισμών tokenization του κειμένου.
- **Νευρωνική συμπίεση ήχου:** Εξέταση της κωδικοποίησης μέσω αλγορίθμων κβαντοποίησης διανυσμάτων (VQ/RVQ).

- **Διατροπική ευθυγράμμιση:** Μελέτη της χαρτογράφησης της προσοχής σε διασταυρούμενα συστήματα.
- **Βελτιστοποίηση και αποκωδικοποίηση:** Ανάλυση των συναρτήσεων απώλειας, της εφαρμογής προηγμένων μεθόδων όπως ο αλγόριθμος Adam και των χρονοπρογραμματιστών δειγματοληψίας (sampling schedulers) για την τελική ανακατασκευή του ακουστικού κύματος κατά τον χρόνο εκτέλεσης (inference).

Κεφάλαιο 4: Συγκριτική Ανάλυση Γενετικών Παραδειγμάτων

Εδώ απομονώνονται και αξιολογούνται τα δύο κύρια αρχιτεκτονικά πλαίσια που διαμορφώνουν σήμερα το τοπίο της παραγωγής μουσικής από κείμενο (text-to-music).

- **Αυτοπαλινδρομική Μοντελοποίηση Γλώσσας (Autoregressive Modeling):** Αντιμετωπίζει τη σύνθεση ήχου ως μια χρονολογική εργασία πρόβλεψης, όπου η παραγωγή γίνεται διαδοχικά, στοιχείο προς στοιχείο (token-by-token).
- **Μοντέλα Λανθάνουσας Διάχυσης (Latent Diffusion Models):** Προσεγγίζουν τη δημιουργία ως μια επαναληπτική διαδικασία αφαίρεσης θορύβου (denoising), η οποία κατευθύνεται από έναν λανθάνοντα χώρο εξαρτώμενο από το κείμενο εισόδου.

Κεφάλαιο 5: Εμπειρικά Σημεία Αναφοράς & Μεθοδολογία Αξιολόγησης

Το κεφάλαιο αυτό ορίζει το αυστηρό μαθηματικό και ποιοτικό πλαίσιο που απαιτείται για τη μέτρηση της απόδοσης των μοντέλων.

- **Αντικειμενικές μετρήσεις:** Εστίαση στην Απόσταση Ήχου Fréchet (FAD) για την αποτίμηση της ακουστικής πιστότητας και στις βαθμολογίες CLAP για τον έλεγχο της σημασιολογικής ευθυγράμμισης κειμένου-ήχου.
- **Υποκειμενική επικύρωση:** Εφαρμογή της Μέσης Βαθμολογίας Γνώμης (MOS) για την ποιοτική αξιολόγηση των αποτελεσμάτων από ανθρώπινους κριτές.

2. Θεωρητικό Υπόβαθρο & Βασικά Στοιχεία

2.1 Επεξεργασία Φυσικής Γλώσσας & Ενσωμάτωση Κειμένου

Κάθε μηχανή διατροφικής σύνθεσης κειμένου σε ήχο βασίζεται σε ένα front-end σύστημα, όπου μετατρέπονται οι μη δομημένες ακολουθίες φυσικής γλώσσας μεταβλητού μήκους σε δομημένες αριθμητικές μορφές. Μόνο τότε μπορεί ένα νευρωνικό δίκτυο να επεξεργαστεί την πληροφορία. Η ροή αυτής της μετατροπής εκτελείται σε δύο διαδοχικά και διακριτά στάδια, τη διακριτή τεμαχιοποίηση (Tokenization) όπου το σύστημα διαχωρίζει τις ακατέργαστες συμβολοσειρές κειμένου σε τοπικά γλωσσικά στοιχεία (tokens) και την ενσωμάτωση διανυσμάτων (Vector Embedding) όπου οι διακριτές αυτές μονάδες αντιστοιχίζονται στη συνέχεια σε έναν πυκνό, συνεχή λανθάνοντα χώρο της μορφής $\mathbb{R}^D$.

2.1.1 Μηχανική Τεμαχιοποίησης

Ιστορικά, τα μοντέλα φυσικής γλώσσας βασίζονταν είτε σε τεμαχιοποίηση σε επίπεδο χαρακτήρων είτε σε επίπεδο λέξεων. Κάθε προσέγγιση, ωστόσο, παρουσίαζε εγγενείς περιορισμούς σε επίπεδο χαρακτήρων, όπου είχε σαν συνέπεια την ακραία επέκταση του μήκους της ακολουθίας, γεγονός που υποβάθμιζε τη μοντελοποίηση εξαρτήσεων μεγάλης εμβέλειας και σε επίπεδο λέξεων όπου δημιουργούσε τεράστια λεξιλόγια και αποτύγχανε πλήρως όταν το σύστημα συναντούσε όρους εκτός λεξιλογίου (Out-of-Vocabulary – OOV). Οι σύγχρονες διατροφικές αρχιτεκτονικές επιλύουν αυτό το πρόβλημα χρησιμοποιώντας αλγόριθμους τεμαχιοποίησης υπολέξεων (subword tokenization). Οι προσεγγίσεις αυτές κατασκευάζουν ένα σταθερό μέγεθος λεξιλογίου V που συνδυάζει χαρακτήρες, ολόκληρες λέξεις και συχνά εμφανιζόμενες μορφηματικές υπολέξεις. Με τον τρόπο αυτό, εξασφαλίζεται ότι οποιαδήποτε περιγραφή —όσο σπάνια κι αν είναι— αναλύεται σε μια έγκυρη ακολουθία γνωστών διακριτικών, αποτρέποντας αστοχίες κατά τον χρόνο εκτέλεσης (inference).

Οι αλγόριθμοι υπολέξεων εξελίχθηκαν από απλά σχήματα καταμέτρησης συχνотήτων σε σύνθετα πιθανοτικά μοντέλα, όπως το BPE ([Sennrich, 2016](#)) όπου ξεκινά από χαρακτήρες και συγχωνεύει επαναληπτικά τα πιο συχνά ζεύγη μέχρι το όριο V , το WordPiece ([Schuster &](#) Διπλωματική Εργασία

[Nakajima, 2012](#)) όπου επιλέγει συγχωνεύσεις βάσει στατιστικής σημασίας, το Unigram ([Kudo, 2018](#)) όπου αντί για άπληστη συγχώνευση, αρχικοποιεί ένα τεράστιο λεξιλόγιο και το κλαδεύει συστηματικά και το BBPE ([Radford et al., 2019](#)) όπου μειώνει το βασικό λεξιλόγιο σε byte UTF-8, εξαλείφοντας πλήρως το διακριτικό άγνωστης λέξης.

Μια σημαντική αρχιτεκτονική διαφοροποίηση εισήχθη με το SentencePiece ([Kudo, 2018](#)). Οι παραδοσιακοί tokenizers απαιτούν εξειδικευμένους προ-tokenizers ανά γλώσσα για να διαχωρίσουν αρχικά το κείμενο με βάση τα κενά διαστήματα. Η διαφορά του SentencePiece είναι ότι αντιμετωπίζει τη συμβολοσειρά εισόδου ως μια ενιαία, συνεχή ροή byte. Χειρίζεται τα κενά διαστήματα ως ξεχωριστούς δομικούς χαρακτήρες (αναπαρίστανται με το σύμβολο `_`), λειτουργώντας ως ένα καθολικό περιτύλιγμα (wrapper) που δεν εξαρτάται από τη γλώσσα και υποστηρίζει τόσο το BPE όσο και το Unigram. Αυτή η προσέγγιση επιτρέπει στους σύγχρονους κωδικοποιητές κειμένου να διαχειρίζονται με ευελιξία τις μη δομημένες προτροπές των χρηστών. Τα ακατέργαστα byte μεταφράζονται άρρηκτα σε πυκνές αναπαραστάσεις, οι οποίες τροφοδοτούν στη συνέχεια τα τελικού σταδίου επίπεδα διασταυρούμενης προσοχής (cross-attention) για την καθοδήγηση της ακουστικής σύνθεσης.

2.1.2 Στατικοί Χώροι Ενσωμάτωσης

Η μετατροπή του κειμένου εισόδου σε μια ακολουθία διακριτών αναγνωριστικών (IDs), $T = \{t_1, t_2, \dots, t_L\}$, αποτελεί το πρώτο βήμα της διαδικασίας. Στη συνέχεια, κάθε διακριτικό πρέπει να αντιστοιχιστεί σε έναν συνεχή διανυσματικό χώρο. Οι πρώτες επιτυχείς προσεγγίσεις βαθιάς μάθησης βασίστηκαν σε στατικές ενσωματώσεις (static embeddings), οι οποίες αναδείχθηκαν μέσα από πλαίσια όπως το Word2Vec ([Mikolov, 2013](#)) και το GloVe ([Pennington, 2014](#)).

Μαθηματικά, μια στατική ενσωμάτωση στηρίζεται σε έναν σταθερό πίνακα αναζήτησης (lookup table), ο οποίος ορίζεται από έναν πίνακα βαρών $E \in \mathbb{R}^{|V| \times D}$. Στο πλαίσιο αυτό, η μεταβλητή $|V|$ αντιπροσωπεύει το μέγεθος του λεξιλογίου και η μεταβλητή D την κρυφή διάσταση (hidden dimension). Για ένα δεδομένο αναγνωριστικό διακριτικού t_i , η

διανυσματική αναπαράσταση e_i εξάγεται μέσω ενός πολλαπλασιασμού : $e_i = x_i E$ όπου το διάνυσμα x_i , με διάσταση $(1 \times |V|)$, περιέχει τη μονάδα αποκλειστικά στον δείκτη t_i .

Το κύριο χαρακτηριστικό των στατικών ενσωματώσεων (*static embeddings*) είναι ότι η συγκεκριμένη αντιστοίχιση των στοιχείων του λεξικού παραμένει εντελώς αμετάβλητη από τα εκάστοτε συμφραζόμενα. Στο επίπεδο της μαθηματικής μοντελοποίησης, ένα διακριτικό t_i (token t_i) αναλύεται πάντα στην ίδια ακριβώς αριθμητική σειρά ενός στατικού πίνακα ενσωματώσεων E , αδιαφορώντας για τη συντακτική δομή ή τη σημασιολογική θέση της λέξης μέσα στην πρόταση. Αν και οι στατικές ενσωματώσεις αποτυπώνουν θεμελιώδεις σημασιολογικές σχέσεις μέσω της διανυσματικής αριθμητικής (*vector arithmetic*) — δημιουργώντας, για παράδειγμα, σαφείς γεωμετρικές γραμμές για μουσικές έννοιες όπως τα είδη ή τα όργανα— πάσχουν από ένα καθοριστικό δομικό πρόβλημα: την πολυσημία (*polysemy*). Όπως τεκμηριώνεται στην κλασική βιβλιογραφία από τους [Peters et al. \(2018\)](#), η αδυναμία αυτή περιορίζει δραστικά τη γενικευτική ικανότητα των εν λόγω αναπαραστάσεων.

Σε έναν στατικό χώρο, η λέξη «μπάσο» αντιστοιχίζεται σε μία και μοναδική συντεταγμένη. Το μοντέλο αδυνατεί δομικά να διακρίνει αν η περιγραφή αφορά την ακουστική ιδιότητα χαμηλών συχνοτήτων ενός τυμπάνου (kick drum) ή ένα είδος ψαριού για παράδειγμα. Αυτή η πλήρης απουσία επίγνωσης του ευρύτερου πλαισίου περιορίζει δραματικά τη χρησιμότητα των στατικών προσεγγίσεων σε συστήματα επεξεργασίας κειμένου σε ήχο (text-to-audio systems). Όπως υπογραμμίζουν οι [Agostinelli et al. \(2023\)](#), στις αρχιτεκτονικές αυτές το τοπικό περιβάλλον ενός επιθέτου ή μιας προσδιοριστικής φράσης μεταβάλλει ριζικά τα φυσικά ακουστικά χαρακτηριστικά που πρέπει τελικά να παραχθούν. Συνεπώς, η στατική αντιμετώπιση των διανυσμάτων αποτυγχάνει να καθοδηγήσει με ακρίβεια τη διαδικασία της ακουστικής σύνθεσης.

2.1.3 Χώροι Ενσωμάτωσης με βάση τα Συμφραζόμενα

Όπως τεκμηριώνεται από τους [Peters et al. \(2018\)](#), οι σύγχρονες διατροφικές αρχιτεκτονικές επιστρατεύουν χώρους ενσωμάτωσης με βάση τα συμφραζόμενα (contextual embedding spaces) για να ξεπεράσουν οριστικά τους περιορισμούς των στατικών αντιστοιχίσεων. Οι

κωδικοποιητές αυτοί δεν εκτελούν μια απλή αναζήτηση διακριτικών (*tokens*) σε έναν δύσκαμπτο πίνακα· αντίθετα, διοχετεύουν ολόκληρη την ακολουθία εισόδου μέσα από βαθιές, πολυεπίπεδες αρχιτεκτονικές προσοχής (*attention architectures*) στη βάση του προτύπου που εισηγήθηκε ο [Vaswani \(2017\)](#). Η προσέγγιση αυτή επιτρέπει στις αριθμητικές τιμές του διανύσματος ενός διακριτικού να μεταβάλλονται δυναμικά, αλληλεπιδρώντας με κάθε άλλο στοιχείο της ακολουθίας.

Η ριζική αλλαγή παραδείγματος σημειώθηκε με μοντέλα όπως το BERT ([Devlin, 2019](#)) και τα αυτοπαλινδρομικά γενετικά πλαίσια ([Radford, 2019](#)). Οι αρχιτεκτονικές αυτές υπολογίζουν έναν πυκνό πίνακα αναπαράστασης $H \in \mathbb{R}^{L \times D}$ από μια ακολουθία εισόδου. Έστω $H^{(0)} = [e_1, e_2, \dots, e_L]^T$ η αρχική στοίβα των στατικών διανυσμάτων εισόδου για μια ακολουθία μήκους L . Ο κωδικοποιητής συμφραζομένων επεξεργάζεται αυτόν τον πίνακα μέσω μιας σειράς M διαδοχικών επιπέδων Transformer. Στο επίπεδο της αρχιτεκτονικής υλοποίησης, η μαθηματική γενίκευση της δυναμικής ενημέρωσης της αναπαράστασης στο επίπεδο m τυποποιείται στη σύγχρονη βιβλιογραφία ως:

$$\mathbf{H}^{(m)} = \text{LayerNorm} \left(\mathbf{H}^{(m-1)} + \text{MultiHeadAttention} \left(\mathbf{H}^{(m-1)} \right) \right) \quad (1)$$

Κατά συνέπεια, το τελικό διάνυσμα ενσωμάτωσης $h_i \in H^{(P)}$ για το διακριτικό t_i παύει να λειτουργεί ως ένας απομονωμένος ορισμός λεξικού. Μετατρέπεται σε μια βαθιά πλασιωμένη τοπολογική συντεταγμένη.

Για παράδειγμα αν στην περίπτωση όπου η προτροπή κειμένου περιέχει τη φράση «βαριά ηλεκτρονική παραμόρφωση μπάσων», τα επίπεδα αυτοπροσοχής συγχωνεύουν μαθηματικά κλάσματα των διανυσμάτων των λέξεων «ηλεκτρονική» και «παραμόρφωση» απευθείας στο διάνυσμα της λέξης «μπάσο» και το αποτέλεσμα είναι η θέση του διακριτικού στον λανθάνοντα χώρο να μετατοπίζεται ρητά. Κάθε βιολογική ή άσχετη ακουστική ασάφεια εξαλείφεται πλήρως. Όταν ο πλασιωμένος αυτός πίνακας μεταβιβάζεται σε ένα σύστημα παραγωγής κειμένου σε μουσική, ο γενετικός πυρήνας αποκτά έναν εξαιρετικά ακριβή και καθαρό σημασιολογικό χάρτη, όπως καταδεικνύεται στα έργα των [Agostinelli et al. \(2023\)](#)

και [Kreuk et al. \(2022\)](#). Ο χάρτης αυτός δεν κωδικοποιεί απλώς τις μεμονωμένες λέξεις. Αποτυπώνει την πλήρη ακουστική, στυλιστική και συναισθηματική πρόθεση του χρήστη.

2.2 Ψηφιακή Επεξεργασία Σήματος (DSP) & Αναπαραστάσεις Ήχου

Η μοντελοποίηση του ήχου μέσω αρχιτεκτονικών βαθιάς μάθησης απαιτεί τη μετατροπή των συνεχών ακουστικών μεταβολών του φυσικού σήματος σε δομημένες, μαθηματικά ανιχνεύσιμες ψηφιακές αναπαραστάσεις. Αυτή η διαδικασία μετάφρασης βασίζεται εξ ολοκλήρου στην Ψηφιακή Επεξεργασία Σήματος (DSP). Κατά την κατασκευή ενός συστήματος μηχανικής μάθησης κειμένου σε ήχο, η θεμελιώδης σχεδιαστική απόφαση αφορά τον χώρο (domain) στον οποίο θα λειτουργήσει το νευρωνικό δίκτυο. Η επιλογή αυτή καθορίζει την υπολογιστική πολυπλοκότητα και την ποιότητα της παραγωγής, και κινείται μεταξύ τριών κύριων αναπαραστάσεων οι οποίες είναι το ακατέργαστο χρονικό πεδίο (Raw Time Domain), το ντετερμινιστικό πεδίο συχνότητας (Deterministic Frequency Domain) και ο ψυχοακουστικά τροποποιημένος λανθάνων χώρος (Mel-Spectrogram).

2.2.1 Το Πεδίο του Χρόνου έναντι του Πεδίου της Συχνότητας

Στην πιο ακατέργαστη μορφή του, ένα ψηφιακό ηχητικό σήμα καταγράφεται στο πεδίο του χρόνου (time domain). Σύμφωνα με την κλασική θεωρία επεξεργασίας σημάτων των [Oppenheim & Schaffer \(2010\)](#), το σήμα αναπαρίσταται ως μια μονοδιάστατη διακριτή ακολουθία τιμών πλάτους (amplitude) της ηχητικής πίεσης, οι οποίες μετρώνται σε ομοιόμορφα χρονικά διαστήματα. Στο επίπεδο της μαθηματικής τυποποίησης, η διακριτή αυτή ακολουθία μοντελοποιείται συμβολικά ως:

$$\mathbf{x} = [x[1], x[2], \dots, x[N]] \quad (2)$$

Ο ρυθμός δειγματοληψίας f_s καθορίζει την πυκνότητα αυτής της ακολουθίας. Η επεξεργασία ακατέργαστων κυματομορφών (raw waveforms) στο πεδίο του χρόνου διατηρεί ανέπαφα όλα

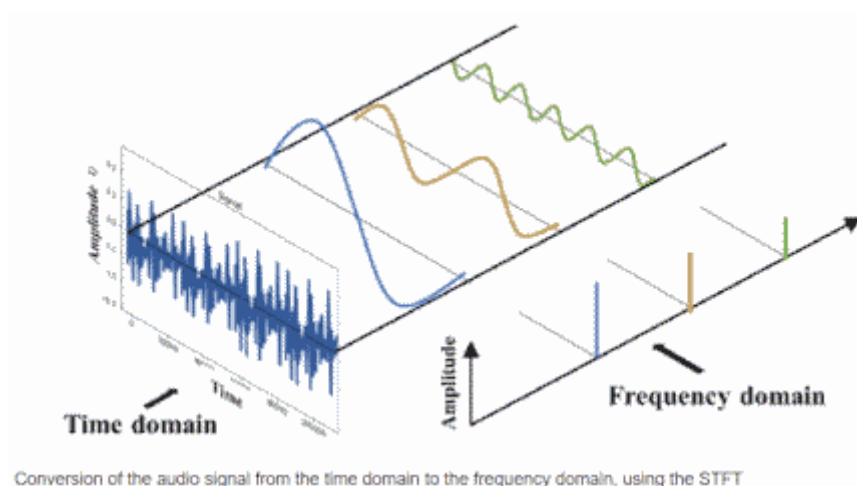
Διπλωματική Εργασία

τα υποκείμενα φυσικά δεδομένα, συμπεριλαμβανομένης της ακριβούς ευθυγράμμισης φάσης (phase alignment) των επικαλυπτόμενων ακουστικών κυμάτων. Ωστόσο, όπως επισημαίνουν οι [van den Oord et al. \(2016\)](#), η επιλογή αυτή δημιουργεί ένα εξαιρετικά απαιτητικό υπολογιστικό περιβάλλον για τα επίπεδα βαθιάς μάθησης (deep learning layers). Το πρόβλημα εντοπίζεται στην ακραία χρονική πυκνότητα (temporal density) και τη χαμηλή σημασιολογική αραιότητα (semantic sparsity) του σήματος. Στο πεδίο του χρόνου, ένα συγκεκριμένο μουσικό χαρακτηριστικό —όπως μια μεμονωμένη νότα πιάνου— δεν περιορίζεται σε έναν μόνο δείκτη. Αντίθετα, διαχέεται σε χιλιάδες υψηλά συσχετισμένες, ταλαντούμενες βαθμωτές μεταβλητές, δυσχεραίνοντας δραματικά την εξαγωγή προτύπων (pattern extraction).

Οι τεχνικές ψηφιακής επεξεργασίας σημάτων (Digital Signal Processing - DSP) μετατοπίζουν το σήμα στο πεδίο της συχνότητας (frequency domain) για να παρακάμψουν αυτή την πυκνότητα δεδομένων. Η μετάβαση αυτή στηρίζεται ιστορικά στη θεωρία Fourier ([Fourier, 1822](#)). Σύμφωνα με τη θεμελιώδη αυτή αρχή, κάθε σύνθετη, συνεχής κυματομορφή μπορεί να αποδομηθεί μαθηματικά σε μια άπειρη υπέρθεση (superposition) βασικών ημιτονοειδών κυμάτων, όπως απεικονίζεται γραφικά στην [Εικόνα 2](#). Στην αριστερή πλευρά του γραφήματος (Πεδίο του Χρόνου - Time domain) αποτυπώνεται το αρχικό, σύνθετο ηχητικό σήμα. Η θεωρία Fourier μάς επιτρέπει να «κοιτάξουμε» στο εσωτερικό του και να το αναλύσουμε στα επιμέρους απλά κύματα που το συνθέτουν (εδώ αναπαρίστανται με το μπλε, το καφέ και το πράσινο κύμα), καθένα από τα οποία έχει διαφορετική συχνότητα και πλάτος. Προβάλλοντας αυτά τα κύματα στη δεξιά πλευρά (Πεδίο της Συχνότητας - Frequency domain), οι συνεχείς ταλαντώσεις μετατρέπονται σε απλές, στατικές κατακόρυφες γραμμές (αιχμές). Η θέση κάθε γραμμής στον οριζόντιο άξονα φανερώνει τη συχνότητά της και το ύψος της το πλάτος (την έντασή) της. Κάθε τέτοιο κύμα ορίζεται μονοσήμαντα από τρεις παραμέτρους: τη συχνότητα, το πλάτος και τη φάση.

Το πλεονέκτημα είναι ότι μεταφέρεται η ανάλυση ήχου στο πεδίο των συχνοτήτων και ένα μοντέλο μηχανικής μάθησης απαλλάσσεται από την ανάγκη να παρακολουθεί ταχείες χρονικές ταλαντώσεις. Αντίθετα, αποκτά τη δυνατότητα να εξετάσει άμεσα τη στατική

παρουσία και την εξέλιξη διακριτών μουσικών τόνων, καθώς και των φασματικών υφών στο πέρασμα του χρόνου. Καθώς κοιτάζουμε το τελικό γράφημα στο πεδίο της συχνότητας (στο βάθος ή στο δεξί μέρος της επεξήγησης), οι χρονικές διακυμάνσεις έχουν πλέον απομονωθεί.



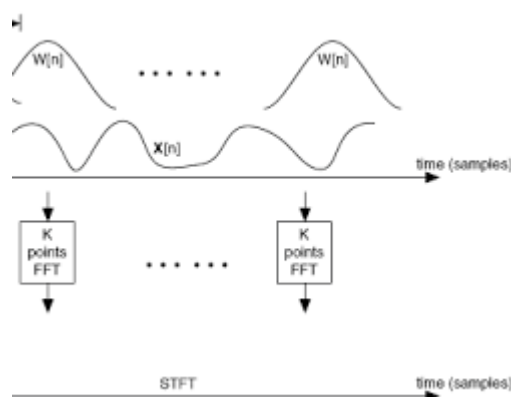
Εικόνα 2: Time Domain to Frequency Domain

2.2.2 Ο Μετασχηματισμός Fourier Βραχύχρονου (STFT)

Κατά την κλασική θεωρία των [Oppenheim & Schaffer \(2010\)](#), ο τυπικός Διακριτός Μετασχηματισμός Fourier (Discrete Fourier Transform - DFT) υπολογίζει τις καθολικές συνιστώσες συχνότητας ενός ολόκληρου σήματος. Με τον τρόπο αυτό, ωστόσο, απορρίπτει πλήρως κάθε χρονικό εντοπισμό της πληροφορίας. Επειδή η μουσική είναι εγγενώς δυναμική —με τις νότες, τις συγχορδίες και τους ρυθμούς να μεταβάλλονται ακαριαία— η εφαρμογή ενός καθολικού DFT αποδεικνύεται δομικά ανεπαρκής. Για τη διατήρηση του χρονικού πλαισίου παράλληλα με τη φασματική σύνθεση, η σύγχρονη βιβλιογραφία βασίζεται στο θεωρητικό πλαίσιο των [Griffin & Lim \(1984\)](#), επιστρατεύοντας τον Μετασχηματισμός Fourier Βραχύχρονου (Short-Time Fourier Transform – STFT).

Ο αλγόριθμος STFT εφαρμόζει μια εντοπισμένη ανάλυση χρόνου-συχνότητας (time-frequency analysis). Όπως τεκμηριώνεται από τον [Harris \(1978\)](#), η μεθοδολογία αυτή στηρίζεται στη σύρση μιας πεπερασμένης συνάρτησης παραθύρου πραγματικών τιμών $w[n]$

—όπως το παράθυρο Hann ή Hamming (Hann or Hamming window)— κατά μήκος του σήματος $x[n]$ στο πεδίο του χρόνου (time domain). Το παράθυρο τεμαχίζει την κυματομορφή σε σύντομα, επικαλυπτόμενα τμήματα (overlapping segments) όπως φαίνεται στην [Εικόνα 3](#) και στη συνέχεια υπολογίζει έναν τοπικό DFT για κάθε ένα από αυτά, επιτρέποντας τη σταδιακή χαρτογράφηση των συχνοτήτων καθώς ο χρόνος εξελίσσεται. Η διαδικασία εκτελείται από πάνω προς τα κάτω: η συνάρτηση παραθύρου (σημειώνεται στο διάγραμμα ως $W[n]$) απομονώνει διαδοχικά τμήματα του σήματος (σημειωμένο ως $X[n]$ στο διάγραμμα). Κάθε απομονωμένο χρονικό παράθυρο τροφοδοτείται ανεξάρτητα σε ένα μπλοκ Γρήγορου Μετασχηματισμού Fourier (αναγράφεται ως « K points FFT», όπου το K ταυτίζεται με το μέγεθος N του αλγορίθμου). Οι έξοδοι αυτοί οι μετασχηματισμών ευθυγραμμίζονται διαδοχικά, δημιουργώντας έναν νέο άξονα frame-by-frame, ο οποίος αποτελεί την τελική αναπαράσταση STFT στο πεδίο χρόνου-συχνότητας.



**Εικόνα 3: Διάγραμμα Μηχανισμού
Ολίσθησης Παραθύρου**

Στο επίπεδο της αρχιτεκτονικής υλοποίησης, ο σύνθετος συντελεστής STFT $X(m, k)$ στο χρονικό πλαίσιο m και για τη ζώνη συχνοτήτων k ορίζεται στη σύγχρονη βιβλιογραφία μέσω της ακόλουθης συμβολικής μοντελοποίησης:

$$X(m, k) = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n - mR] \cdot e^{-j \frac{2\pi}{N} k n} \quad (3)$$

Οι βασικές παράμετροι της εξίσωσης ορίζονται ως εξής:

- R (Hop Size): Το μέγεθος του άλματος (hop size), δηλαδή ο αριθμός των δειγμάτων που προχωρά η συνάρτηση παραθύρου σε κάθε υπολογιστικό βήμα. .
- N (FFT Size): Το μέγεθος του Γρήγορου Μετασχηματισμού Fourier, το οποίο καθορίζει τον συνολικό αριθμό των διακριτών ζωνών συχνότητας.

Η έξοδος του STFT είναι ένας πίνακας σύνθετων αριθμών $X(m, k) \in \mathbb{C}$, ο οποίος αναλύεται πολικά σε δύο διακριτά στοιχεία:

1. Φασματογράφημα Μεγέθους $|X(m, k)|$: Αντιπροσωπεύει την απόλυτη ενέργεια ή την ένταση της συχνότητας k στο χρονικό πλαίσιο m .
2. Φασματογράφημα Φάσης $\angle X(m, k)$: Αποτυπώνει την εντοπισμένη ευθυγράμμιση κύματος (wave alignment) της συχνότητας k στο χρονικό πλαίσιο m .

Η χρήση του STFT εισάγει έναν ανυπέρβλητο φυσικό περιορισμό που πηγάζει από την Αρχή Αβεβαιότητας του Heisenberg για την ανάλυση χρόνου-συχνότητας όπως αυτή θεμελιώθηκε από τον [Gabor \(1946\)](#). Στη σύγχρονη ψηφιακή επεξεργασία σημάτων, η επιλογή του μεγέθους του παραθύρου N επιβάλλει έναν απόλυτο συμβιβασμό (resolution trade-off):

- Ευρύ παράθυρο ($N \uparrow$): Προσφέρει εξαιρετική ανάλυση συχνότητας (frequency resolution), επιτρέποντας τον λεπτομερή διαχωρισμό του μουσικού ύψους, αλλά υποβαθμίζει τη χρονική ανάλυση (temporal resolution), θολώνοντας τους ακριβείς χρόνους έναρξης των φθόγων.
- Στενό παράθυρο ($N \downarrow$): Δίνει προτεραιότητα στον ακριβή χρονικό εντοπισμό των ακουστικών γεγονότων, προκαλεί όμως έντονη φασματική διαρροή (spectral leakage), περιορίζοντας τη φασματική ευκρίνεια.

Στο επίπεδο της αρχιτεκτονικής υλοποίησης των σύγχρονων δικτύων, τα περισσότερα μοντέλα μηχανικής μάθησης απορρίπτουν το χαοτικό φάσμα φάσης $\angle X(m, k)$ για να επικεντρωθούν αποκλειστικά στη δημιουργία του ιδιαίτερα δομημένου φασματογράμματος

μεγέθους $|X(m, k)|$. Όπως αναλύουν οι [Griffin & Lim \(1984\)](#), αυτή η σχεδιαστική επιλογή μετατρέπει την ανακατασκευή του τελικού ακουστικού κύματος σε ένα κακώς ορισμένο πρόβλημα αντιστροφής (ill-posed inversion problem). Ιστορικά, η επίλυση αυτής της μαθηματικής ασυνέχειας απαιτούσε τη χρήση επαναληπτικών αλγορίθμων εκτίμησης της χαμένης φάσης —με χαρακτηριστικότερο παράδειγμα τον ομώνυμο αλγόριθμο Griffin-Lim— προτού καταστεί δυνατή η σύνθεση μιας αναπαραγωγίσιμης και καθαρής κυματομορφής στο πεδίο του χρόνου. Στη σύγχρονη βιβλιογραφία, η ανάγκη για τέτοιες επαναληπτικές προσεγγίσεις έχει οδηγήσει στην ανάπτυξη προηγμένων νευρωνικών κωδικοποιητών φωνής (neural vocoders) που επιχειρούν να παρακάμψουν αυτόν τον περιορισμό.

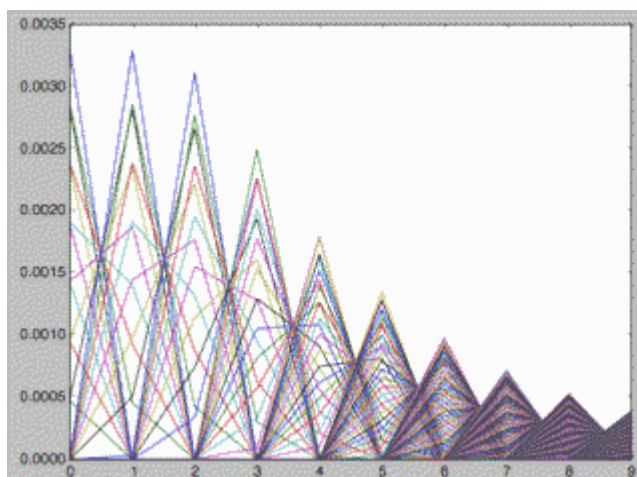
2.2.3 Φασματογράμματα Mel και Λογαριθμική Συμπίεση

Το φασματογράφημα μεγέθους STFT παρέχει μια πλήρη μαθηματική αναπαράσταση των ακουστικών συχνοτήτων. Ωστόσο, δεν είναι βελτιστοποιημένο για την ανθρώπινη ακουστική αντίληψη ούτε για την εξαγωγή χαρακτηριστικών (feature extraction) από μοντέλα βαθιάς μάθησης (deep learning). Όπως τεκμηριώνεται από τους [Zwicker & Fastl \(1999\)](#), η ανθρώπινη ακοή δεν αντιλαμβάνεται τη συχνότητα ή την ένταση σε γραμμική κλίμακα. Το ακουστικό μας σύστημα εμφανίζει εξαιρετική ευαισθησία σε μικρές διακυμάνσεις ύψους στις χαμηλές συχνότητες (διακρίνοντας εύκολα τη διαφορά μεταξύ 200 Hz και 300 Hz, αλλά παραμένει σε μεγάλο βαθμό αναισθητό σε πανομοιότυπες γραμμικές μεταβολές στα ανώτερα μητρώα (για παράδειγμα, η διαφορά μεταξύ 14.000 Hz και 14.100 Hz είναι πρακτικά μη αντιληπτή).

Για να γεφυρωθεί αυτό το αντιληπτικό χάσμα, οι γραμμικές συχνότητες του STFT αντιστοιχίζονται στην κλίμακα Mel (Mel scale), η οποία αναπτύχθηκε αρχικά από τον [Stevens \(1937\)](#). Πρόκειται για μια μη γραμμική, υποκειμενική κλίμακα ακουστικού ύψους (pitch), σχεδιασμένη να προσομοιώνει το ανθρώπινο ψυχοακουστικό προφίλ. Στη σύγχρονη βιβλιογραφία, η μαθηματική μετατροπή από τη φυσική συχνότητα f (σε Hertz) στο αντιληπτικό ύψος m (σε Mels) τυποποιείται από τη σχέση:

$$m = 2595 \cdot \log_{10} \left(1 + \frac{f}{700} \right) \quad (4)$$

Για τον υπολογισμό του Φασματογραφήματος Mel, ο πίνακας γραμμικού μεγέθους $|X(m, k)|$ πολλαπλασιάζεται με μια τράπεζα φίλτρων Mel (Mel filterbank). Η δομή αυτής της τράπεζας αποτελείται από επικαλυπτόμενα τριγωνικά φίλτρα ζώνης διέλευσης (triangular bandpass filters). Τα φίλτρα αυτά είναι κατανομημένα γραμμικά κατά μήκος της κλίμακας Mel, γεγονός που σημαίνει ότι στο φυσικό πεδίο των Hertz γίνονται προοδευτικά ευρύτερα και πιο συμπιεσμένα όσο αυξάνεται η συχνότητα, όπως φαίνεται στην [Εικόνα 4](#) που αποτυπώνεται η τράπεζα φίλτρων Mel πάνω σε έναν γραμμικό άξονα συχνοτήτων. Παρατηρούμε ότι καθώς κινούμαστε προς τις υψηλές συχνότητες (δεξιά), οι βάσεις των τριγώνων διευρύνονται (κάθε φίλτρο καλύπτει μεγαλύτερο εύρος συχνοτήτων σε Hz), ενώ το ύψος τους μειώνεται σταδιακά (κανονικοποίηση ενέργειας), αποτυπώνοντας έτσι τη μειωμένη διακριτική ικανότητα της ανθρώπινης ακοής στα πρίμα.



Εικόνα 4: Mel filterbank triangular filters diagram

Παράλληλα, η ανθρώπινη αντίληψη του ακουστικού όγκου ακολουθεί λογαριθμική συμπεριφορά βάσει του νόμου Weber-Fechner, όπως επισημαίνουν οι προαναφερθέντες [Zwicker & Fastl \(1999\)](#). Προκειμένου να ευθυγραμμιστούν οι τιμές ακατέργαστης ενέργειας

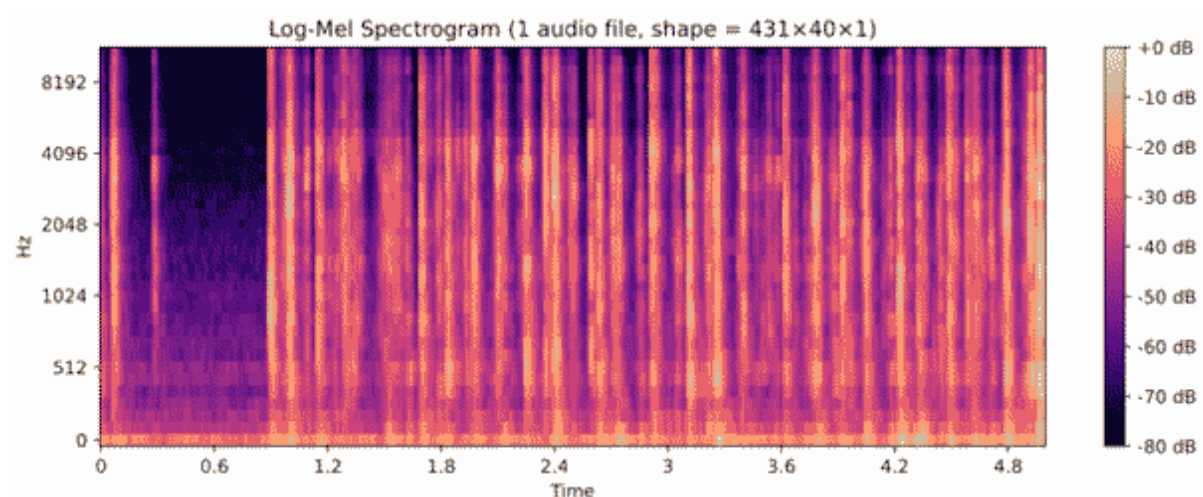
του φασματογραφήματος με την ανθρώπινη αίσθηση της έντασης, τα απόλυτα πλάτη κλιμακώνονται μέσω ενός λογαριθμικού συμπίεστη (logarithmic compressor). Σύμφωνα με το μεθοδολογικό πλαίσιο που εισήγαγε ο [Wang \(2017\)](#), στο επίπεδο της αρχιτεκτονικής υλοποίησης η διαδικασία αυτή για την παραγωγή του Φασματογραφήματος Log-Mel (Log-Mel spectrogram) μπορεί να αποδοθεί συμβολικά ως:

$$S_{Mel}(m, \phi) = \log(|X(m, k)| \cdot M_{filterbank} + \epsilon) \quad (5)$$

όπου η μεταβλητή ϵ αποτελεί μια απειροελάχιστη θετική σταθερά που αποτρέπει τον υπολογισμό του λογαρίθμου του μηδενός.

Στο πλαίσιο των σύγχρονων διατροπικών αρχιτεκτονικών (cross-modal architectures), ο [Wang \(2017\)](#) επισημαίνει ότι το Φασματογράφημα Log-Mel λειτουργεί ως μια εξαιρετικά αποτελεσματική, στατική δομική γέφυρα. Συμπυκνώνοντας εκατομμύρια ακατέργαστα δείγματα του πεδίου του χρόνου σε έναν πίνακα συντεταγμένων χαμηλής διάστασης, το μοντέλο μορφοποιεί τα ηχητικά δεδομένα ακριβώς όπως μια ψηφιακή εικόνα ενός καναλιού, η οποία εμφανίζεται στην [Εικόνα 5](#). Όπως φαίνεται στο συγκεκριμένο παράδειγμα, το φασματογράφημα Log-Mel αποκτά μια ξεκάθαρη δισδιάστατη δομή (εδώ με διαστάσεις $431 \times 40 \times 1$):

- Ο οριζόντιος άξονας (Time): Αντιπροσωπεύει τη χρονική εξέλιξη (στην προκειμένη περίπτωση, μια διάρκεια περίπου 5 δευτερολέπτων χωρισμένη σε 431 χρονικά frames).
- Ο κατακόρυφος άξονας (Hz): Αντιπροσωπεύει το αντιληπτικό ύψος, κατανεμημένο μη γραμμικά σε 40 ζώνες συχνοτήτων Mel (από 0 έως 8192 Hz).
- Οι τιμές των εικονοστοιχείων (Pixels): Αναπαριστούν τη λογαριθμικά συμπίεσμένη ένταση (ενέργεια) του σήματος. Η χρωματική κλίμακα στα δεξιά δείχνει ότι τα φωτεινά/κίτρινα σημεία αντιστοιχούν σε υψηλή ακουστική ενέργεια (κοντά στα 0 dB), ενώ τα σκοτεινά/μωβ σημεία προσεγγίζουν τη σιγή (-80 dB).



Εικόνα 5:Log-Mel Spectrogram

Αυτός ο μετασχηματισμός προσφέρει δύο καθοριστικά πλεονεκτήματα:

- Απομακρύνει τον χαοτικό, μη ντετερμινιστικό θόρυβο φάσης(phase noise), ο οποίος δυσχεραίνει τη σταθερή εκπαίδευση των δικτύων.
- Διατηρεί ανέπαφα τα πυκνά αρμονικά πρότυπα (harmonic patterns), τα ηχοχρώματα (timbres) των διαφορετικών οργάνων και τις δομικές διατάξεις των μουσικών φθόγγων.

Αυτή η χωρική μορφοποίηση επιτρέπει στα γενετικά μοντέλα (generative models) να αξιοποιούν βελτιστοποιημένα δισδιάστατα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs) και επίπεδα αυτοπροσοχής πολλαπλών κεφαλών (Multi-Head Self-Attention layers). Με τον τρόπο αυτό, όπως καταδεικνύεται από τους [Kong et al. \(2020\)](#), τα δίκτυα συνθέτουν ολοκληρωμένες μουσικές δομές προτού αυτές μετατραπούν ξανά, μέσω κατάλληλων νευρωνικών αποκωδικοποιητών φωνής (neural vocoders), σε ακατέργαστες ακουστικές κυματομορφές (raw waveforms).

2.3 Αρχιτεκτονικές Βαθιάς Μάθησης για Χαρτογράφηση Ακολουθιών

Η διατροφική παραγωγή ήχου από κείμενο αποτελεί στη βάση του ένα πρόβλημα χαρτογράφησης ακολουθίας σε ακολουθία (sequence-to-sequence), όπου διανύσματα κειμενικών χαρακτηριστικών μεταφράζονται σε μια ευθυγραμμισμένη σειρά ακουστικών αναπαραστάσεων. Η αρχιτεκτονική εξέλιξη ορίζεται από τη μετατόπιση από τα επαναληπτικά δίκτυα, τα οποία επεξεργάζονται τα δεδομένα βήμα-προς-βήμα δημιουργώντας ένα στενό χρονικό σημείο συμφόρησης, στα πλαίσια προσοχής Transformer, μηχανισμοί οι οποίοι καταργούν τον διαδοχικό περιορισμό, μοντελοποιούν τις καθολικές σχέσεις της ακολουθίας παράλληλα και εισάγουν εξειδικευμένες δομικές προσαρμογές, βελτιστοποιημένες ειδικά για τη διατήρηση της ακουστικής και μουσικής συνοχής σε μεγάλες χρονικές κλίμακες.

2.3.1 Επαναληπτικές Προσεγγίσεις: RNN, LSTM και το Sequential Bottleneck

1)Επαναλαμβανόμενο νευρωνικό δίκτυο (RNN)

Σύμφωνα με το θεμελιώδες έργο του [Elman \(1990\)](#), τα πρώτα επιτυχημένα πλαίσια για τη μοντελοποίηση ακολουθιών δεδομένων βασίστηκαν στα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNN). Το καθοριστικό δομικό χαρακτηριστικό αυτής της αρχιτεκτονικής είναι ο επαναλαμβανόμενος βρόχος ανάδρασης, ο οποίος επιτρέπει τη διατήρηση και τη ροή της πληροφορίας κατά μήκος των διαδοχικών χρονικών βημάτων. Στη σύγχρονη βιβλιογραφία, η μαθηματική γενίκευση ενός βασικού (vanilla) RNN τυποποιείται ως εξής: δεδομένης μιας ακολουθίας εισόδου $X = [x_1, x_2, \dots, x_T]$, το δίκτυο υπολογίζει ένα διάνυσμα κρυφής κατάστασης (hidden state) h_t σε κάθε χρονικό βήμα t . Αυτό επιτυγχάνεται συνδυάζοντας το τρέχον διάνυσμα χαρακτηριστικών εισόδου με το αμέσως προηγούμενο διάνυσμα κρυφής κατάστασης:

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t + b_h) \quad (6)$$

Οι παράμετροι της διανυσματικής αυτής μοντελοποίησης ορίζονται ως εξής:

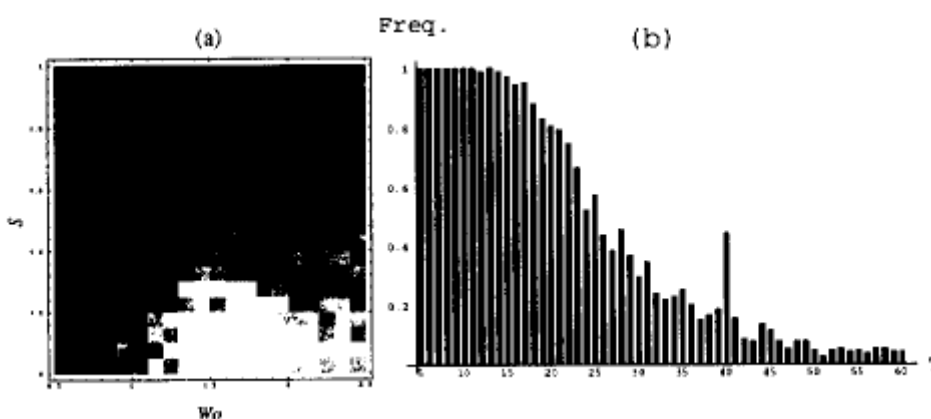
- W_{hh} : Ο επαναλαμβανόμενος πίνακας βαρών (recurrent weight matrix) που μεταφέρει την πληροφορία από την προηγούμενη κρυφή κατάσταση.
- W_{xh} : Ο πίνακας βαρών εισόδου (input weight matrix) που προβάλλει το τρέχον στοιχείο της ακολουθίας στον κρυφό χώρο.
- b_h : Το διάνυσμα πόλωσης (bias vector).

Αυτός ο επαναλαμβανόμενος υπολογιστικός κύκλος διαμορφώνει μια δυναμική εσωτερική μνήμη. Το δίκτυο παύει να εξετάζει τα στοιχεία της εισόδου απομονωμένα, αποκτώντας τη δυνατότητα αναγνώρισης μακροπρόθεσμων εξαρτήσεων. Αντίθετα, η συγκεκριμένη προσέγγιση παρακολουθεί ενεργά το προηγούμενο χρονικό πλαίσιο. Όπως επισημαίνει ο δημιουργός, η ιδιότητα αυτή επιτρέπει στο μοντέλο να αναγνωρίζει με ακρίβεια τη χρονική αλληλουχία, ενσωματώνοντας αποτελεσματικά το ευρύτερο σημασιολογικό και συντακτικό πλαίσιο της επεξεργαζόμενης ακολουθίας.

2) Πρόβλημα των κλίσεων και Δίκτυο Μακράς Βραχυπρόθεσμης Μνήμης (LSTM)

Τα τυπικά (vanilla) RNN υποφέρουν από έναν σοβαρό μαθηματικό περιορισμό: το πρόβλημα της εξαφάνισης (vanishing) και της έκρηξης (exploding) των διαβαθμίσεων. Όπως τεκμηριώνεται στην καθοριστική μελέτη των [Bengio et al. \(1994\)](#), όταν η αντίστροφη διάδοση (backpropagation) υπολογίζεται μέσω μεγάλων χρονικών οριζόντων, οι κλίσεις σφάλματος πολλαπλασιάζονται επανειλημμένα με τον επαναλαμβανόμενο πίνακα βαρών W_{hh} . Εάν οι ιδιοτιμές (eigenvalues) αυτού του πίνακα είναι μικρότερες από τη μονάδα, το σήμα του σφάλματος μειώνεται εκθετικά προσεγγίζοντας το μηδέν. Το φαινόμενο αυτό εμποδίζει δομικά το μοντέλο από το να εκπαιδευτεί και να συγκρατήσει μακροπρόθεσμες εξαρτήσεις, όπως απεικονίζεται στην [Εικόνα 6](#). Πιο συγκεκριμένα στην εικόνα (a), εξερευνάται ο χώρος των παραμέτρων του δικτύου, όπου η εκτεταμένη σκοτεινή περιοχή αντιπροσωπεύει τις συνθήκες όπου η κλίση είτε μηδενίζεται (vanishing) είτε εκτινάσσεται (exploding), αφήνοντας μόνο ένα ελάχιστο «παράθυρο» σταθερής εκπαίδευσης (λευκά στοιχεία). Αντίστοιχα, στο ιστορικό γράφημα (b), απεικονίζεται η συχνότητα επιτυχούς

σύγκλισης (Freq.) του vanilla RNN σε σχέση με το μέγεθος της χρονικής υστέρησης T (time steps). Γίνεται άμεσα αντιληπτό ότι ενώ για μικρούς χρονικούς ορίζοντες $T < 10$ το δίκτυο αποδίδει τέλεια, η υπολογιστική του ικανότητα να συγκρατήσει πληροφορίες κατακρημνίζεται καθώς το T αυξάνεται, αποδεικνύοντας πειραματικά την αδυναμία διατήρησης μακροπρόθεσμης μνήμης.



Εικόνα 6 :Πειραματικά αποτελέσματα σύγκλισης της εκπαίδευσης σε απλά RNN

Για την άρση αυτής της αρχιτεκτονικής αδυναμίας, οι [Hochreiter & Schmidhuber \(1997\)](#) εισήγαγαν το δίκτυο Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory - LSTM). Η προσέγγιση αυτή αντικαθιστά τον απλό επαναλαμβανόμενο νευρώνα με μια σύνθετη δομή κυτταρικής μνήμης (cell memory block). Η ροή και ο δυναμικός έλεγχος της πληροφορίας στο εσωτερικό της δομής διέπονται από τρεις μαθηματικές πύλες:

- Πύλη Λήθης (f_t - Forget Gate): Καθορίζει ποιο ποσοστό της προηγούμενης γνώσης θα απορριφθεί, μια δομική προσθήκη που σταθεροποιήθηκε βιβλιογραφικά από τους [Gers et al. \(2000\)](#).
- Πύλη Εισόδου (i_t - Input Gate): Ρυθμίζει ποιες νέες πληροφορίες από την τρέχουσα είσοδο θα αποθηκευτούν στην κατάσταση του κελιού.

- Πύλη Εξόδου (o_t - Output Gate): Αποφασίζει ποιο τμήμα της εσωτερικής μνήμης θα περάσει στην κρυφή κατάσταση (hidden state) ως έξοδος του τρέχοντος βήματος.

Στο επίπεδο της αρχιτεκτονικής υλοποίησης, το σύστημα των εξισώσεων που περιγράφει τη δυναμική ανανέωσης των εσωτερικών μεταβλητών ενός κελιού LSTM σε κάθε χρονικό βήμα t τυποποιείται στη σύγχρονη βιβλιογραφία μέσω της ακόλουθης συμβολικής μοντελοποίησης:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (7)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (8)$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (9)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (10)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (11)$$

$$h_t = o_t \odot \tanh(C_t) \quad (12)$$

Όπου:

- σ : Αντιπροσωπεύει τη μη γραμμική συνάρτηση ενεργοποίησης σιγμοειδούς (sigmoid function).
- $\odot$: Δηλώνει τον πολλαπλασιασμό κατά στοιχεία (element-wise / Hadamard product).
- C_t : Είναι η τρέχουσα κατάσταση του κελιού (cell state).

Η κατάσταση του κελιού C_t λειτουργεί ως ένας εσωτερικός «μεταφορικός ιμάντας» πληροφορίας. Οι δημιουργοί επισημαίνουν ότι η συγκεκριμένη διάταξη επιτρέπει στις διαβαθμίσεις του σφάλματος να ρέουν ανεμπόδιστα προς τα πίσω στον χρόνο μέσω απλών γραμμικών προσθέσεων. Με τον τρόπο αυτό, παρακάμπτονται πλήρως οι διαδοχικοί, καταστροφικοί πολλαπλασιασμοί πινάκων που προκαλούσαν την εξαφάνιση των κλίσεων στα συμβατικά επαναλαμβανόμενα δίκτυα.

3) Το διαδοχικό σημείο συμφόρησης(Sequential Bottleneck)

Παρά την εισαγωγή των μηχανισμών πύλης (gates) στα LSTM, οι επαναλαμβανόμενες αρχιτεκτονικές παρουσιάζουν ένα θεμελιώδες δομικό και υπολογιστικό ελάττωμα. Το φαινόμενο αυτό είναι γνωστό στη βιβλιογραφία ως διαδοχικό σημείο συμφόρησης (sequential bottleneck) και, όπως επισημαίνουν οι [Vaswani et al. \(2017\)](#), περιορίζει δραματικά την αποδοτικότητά τους σε μεγάλης κλίμακας προβλήματα.

Οι αδυναμίες των RNN και LSTM εντοπίζονται σε δύο άξονες: την υπολογιστική τους εξάρτηση από τον χρόνο και την πεπερασμένη χωρητικότητα της μνήμης τους.

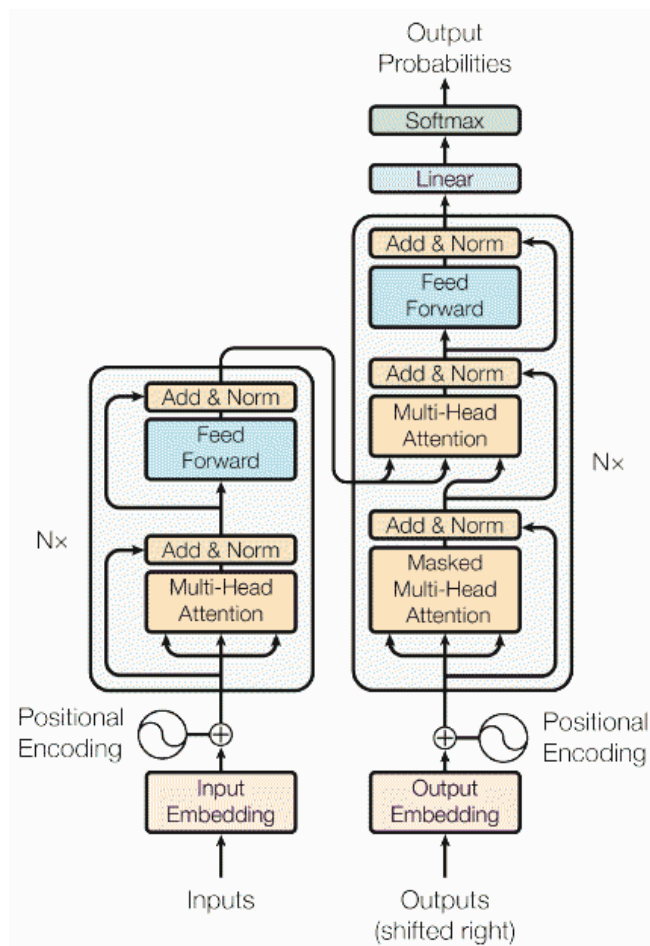
- Έλλειψη Παραλληλοποίησης (Lack of Parallelization): Στο επίπεδο της μαθηματικής μοντελοποίησης, καθώς ο υπολογισμός της τρέχουσας κρυφής κατάστασης h_t προϋποθέτει τη γνώση της αμέσως προηγούμενης κατάστασης h_{t-1} στη βάση του ιστορικού προτύπου του [Elman \(1990\)](#), η εκπαίδευση του δικτύου είναι αδύνατο να παραλληλιστεί κατά μήκος της χρονικής διάστασης (temporal dimension). Αν και το σύγχρονο υλικό (GPU) είναι σχεδιασμένο να εκτελεί χιλιάδες υπολογισμούς πινάκων ταυτόχρονα, η επαναλαμβανόμενη φύση των RNN αναγκάζει τους επεξεργαστές να λειτουργούν γραμμικά, βήμα προς βήμα. Αυτό το υπολογιστικό εμπόδιο δημιουργεί σοβαρά προβλήματα καθυστέρησης κατά την επεξεργασία μεγάλων ηχητικών ακολουθιών.
- Απώλεια Πληροφοριών σε Μακρούς Ορίζοντες (Information Loss over Long Horizons): Το δίκτυο επιχειρεί να συμπίεσει μια διαρκώς επεκτεινόμενη ιστορική χρονογραμμή σε ένα μόνο διάνυσμα σταθερού μεγέθους h_t . Καθώς το μήκος της ακολουθίας T αυξάνεται σε χιλιάδες βήματα —κάτι εξαιρετικά σύνηθες σε μεγάλα κείμενα ή σύνθετα ακουστικά σήματα— οι παλαιότερες πληροφορίες αντικαθίστανται και εξασθενούν προοδευτικά από τα νέα δεδομένα εισόδου. Το φαινόμενο αυτό, σύμφωνα με τους [Bengio et al. \(1994\)](#) και [Hochreiter & Schmidhuber \(1997\)](#), έχει ως αποτέλεσμα την απώλεια της δομικής συνέπειας και της μακροπρόθεσμης συνοχής του συστήματος.

Αυτό το διπλό ελάττωμα κατέστησε σαφές ότι για την επιτυχή διατροπική σύνθεση κειμένου σε ήχο (text-to-audio synthesis) υψηλής πιστότητας, η επιστήμη της μηχανικής μάθησης (machine learning) έπρεπε να εγκαταλείψει την επεξεργασία με βάση τη χρονολογική σειρά και να στραφεί σε αρχιτεκτονικές που βασίζονται στην καθολική, παράλληλη προσοχή (global, parallel attention).

2.3.2 Η Αρχιτεκτονική του Transformer και η Μηχανική της Αυτοπροσοχής

Το διαδοχικό σημείο συμφόρησης των επαναλαμβανόμενων δικτύων παρακάμφθηκε οριστικά με την εισαγωγή της αρχιτεκτονικής Transformer από τους [Vaswani et al. \(2017\)](#). Ο Transformer απορρίπτει πλήρως τους επαναλαμβανόμενους βρόχους ανάδρασης, καθώς βασίζεται αποκλειστικά σε έναν μαθηματικό μηχανισμό που ονομάζεται αυτοπροσοχή (self-attention). Ο μηχανισμός αυτός επιτρέπει τη χαρτογράφηση των εξαρτήσεων μιας ακολουθίας ταυτόχρονα, αδιαφορώντας για τη χρονολογική απόσταση των στοιχείων της.

Όπως φαίνεται στην [Εικόνα 7](#), ο Transformer ακολουθεί μια δομή Κωδικοποιητή-Αποκωδικοποιητή (Encoder-Decoder). Το αριστερό τμήμα (Encoder) επεξεργάζεται την ακολουθία εισόδου, ενώ το δεξί (Decoder) παράγει την ακολουθία εξόδου με αυτοπαλινδρομικό τρόπο. Κάθε ένα από αυτά τα δύο μέρη αποτελείται από μια στοίβα N πανομοιότυπων επιπέδων. Πέρα από τον κεντρικό μηχανισμό της Προσοχής Πολλαπλών Κεφαλών (Multi-Head Attention), κάθε επίπεδο περιλαμβάνει ένα πλήρως συνδεδεμένο δίκτυο πρόσθιας τροφοδότησης (Feed Forward), το οποίο εφαρμόζεται σε κάθε θέση ξεχωριστά. Για τη διασφάλιση της σταθερότητας κατά την εκπαίδευση, χρησιμοποιούνται υπολειμματικές συνδέσεις (residual connections) γύρω από κάθε υπο-επίπεδο, ακολουθούμενες από κανονικοποίηση επιπέδου (Layer Normalization), μια διαδικασία που αποτυπώνεται ως «Add & Norm». Τέλος, επειδή το μοντέλο στερείται επαναληπτικότητας, η πληροφορία για τη σχετική σειρά των στοιχείων εισάγεται στη βάση της αρχιτεκτονικής μέσω των διανυσμάτων Θεσιακής Κωδικοποίησης (Positional Encoding).



Εικόνα 7: Transformer-Model Architecture

Αντί να επεξεργάζεται τις εισόδους διακριτικό προς διακριτικό, η αρχιτεκτονική αυτή προβάλλει ολόκληρο τον πίνακα αναπαράστασης εισόδου $H \in \mathbb{R}^{T \times D}$ σε τρεις διακριτούς χώρους χρησιμοποιώντας εκπαιδευσίμους πίνακες γραμμικής προβολής: τα Ερωτήματα (Queries - Q), τα Κλειδιά (Keys - K) και τις Τιμές (Values - V). Στο επίπεδο της αρχιτεκτονικής υλοποίησης, οι γραμμικές αυτές προβολές μοντελοποιούνται συμβολικά ως:

$$Q = H W_Q \quad K = H W_K \quad V = H W_V \quad (14)$$

$$W_Q, W_K \in \mathbb{R}^{D \times d_k} \quad \text{και} \quad W_V \in \mathbb{R}^{D \times d_v} \quad (15)$$

Το σύστημα υπολογίζει το επίπεδο εστίασης κάθε μεμονωμένου διακριτικού προς όλα τα υπόλοιπα στοιχεία της ακολουθίας, εξάγοντας το εσωτερικό γινόμενο του πίνακα ερωτήσεων με την αναστροφή του πίνακα κλειδιών QK^T .

Σε υψηλές διαστάσεις, οι τιμές του εσωτερικού γινομένου τείνουν να αυξάνονται υπερβολικά, ωθώντας τη συνάρτηση softmax σε περιοχές με επικίνδυνα μικρές κλίσεις (vanishing gradients). Για να αποτραπεί αυτό, οι βαθμολογίες κλιμακώνονται με την τετραγωνική ρίζα της διάστασης των κλειδιών d_k .

Η μαθηματική γενίκευση της κλιμακωτής προσοχής εσωτερικού γινομένου (scaled dot-product attention) τυποποιείται ως:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (16)$$

Η συνάρτηση softmax παράγει έναν πίνακα βαρών προσοχής διάστασης $T \times T$, όπου κάθε γραμμή αποτελεί μια κανονικοποιημένη κατανομή πιθανοτήτων. Αυτός ο πίνακας καθορίζει με ακρίβεια το ποσοστό των πληροφοριών περιβάλλοντος από τον πίνακα τιμών $\mathbf{V}$ που θα ενσωματωθεί στην τελική αναπαράσταση κάθε διακριτικού.

Προκειμένου το μοντέλο να μπορεί να παρακολουθεί ταυτόχρονα διαφορετικούς τύπους σχέσεων και πληροφοριών, το πλαίσιο αυτό επιστρατεύει την προσοχή πολλαπλών κεφαλών (multi-head attention). Αντί για έναν μοναδικό υπολογισμό προσοχής σε ολόκληρη την κρυφή διάσταση, τα ερωτήματα, τα κλειδιά και οι τιμές τεμαχίζονται σε h ανεξάρτητους υποχώρους. Κάθε «κεφαλή»(head) εκτελεί παράλληλα τη δική της κλιμακωτή συνάρτηση προσοχής. Στη σύγχρονη βιβλιογραφία, η διαδικασία αυτή στο επίπεδο της αρχιτεκτονικής υλοποίησης τυποποιείται μέσω του ακόλουθου συστήματος συμβολικής μοντελοποίησης:

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}^O \quad (17)$$

όπου

$$\text{head}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_Q^i, \mathbf{K}\mathbf{W}_K^i, \mathbf{V}\mathbf{W}_V^i\right) \quad (18)$$

Στην πλευρά του αποκωδικοποιητή ([Εικόνα 7](#)), ο μηχανισμός αυτός τροποποιείται σε Masked Multi-Head Attention, εμποδίζοντας το μοντέλο να έχει πρόσβαση σε μελλοντικές θέσεις κατά τη διάρκεια της παραγωγής κειμένου ή μουσικής. Αυτή η πολυεπίπεδη προβολή προσφέρει εξαιρετική αναπαραστατική ευελιξία. Επιτρέπει, για παράδειγμα, σε μια κεφαλή προσοχής να εστιάζει στις άμεσες συντακτικές συσχετίσεις των λέξεων μέσα σε μια προτροπή κειμένου, την ίδια στιγμή που μια γειτονική κεφαλή παρακολουθεί παράλληλα ευρύτερες δομικές έννοιες, όπως το μουσικό είδος, το επιθυμητό τέμπο ή τις ρυθμικές δεσμεύσεις της ακουστικής σύνθεσης.

2.3.3 Τοπολογίες Προσοχής: Αιτιώδης Διαμόρφωση Μάσκας έναντι Αμφίδρομου Πλαισίου

Η διαμόρφωση της μάσκας (masking) στον πίνακα προσοχής (attention matrix) αποτελεί μια κρίσιμη αρχιτεκτονική παραλλαγή για τη σωστή χαρτογράφηση των ακολουθιών, καθώς καθορίζει τον τρόπο και την κατεύθυνση με την οποία θα ρεύσει η πληροφορία μέσα στο νευρωνικό δίκτυο.

1) Αμφίδρομες Τοπολογίες (Bidirectional Topologies)

Στις μη αιτιώδεις ή αμφίδρομες τοπολογίες (non-causal or bidirectional topologies), ο πίνακας προσοχής επιτρέπει την καθολική και χωρίς περιορισμούς δρομολόγηση των πληροφοριών. Όπως τεκμηριώνεται από τους [Devlin et al. \(2019\)](#), οι δομές αυτές συναντώνται συχνά σε κωδικοποιητές κειμένου (text encoders) όπως το BERT. Παράλληλα, αξιοποιούνται στα χωροχρονικά μπλοκ των δικτύων διάχυσης (spatiotemporal blocks of diffusion models), σύμφωνα με το μεθοδολογικό πλαίσιο των [Ho et al. \(2020\)](#). Στο εσωτερικό αυτού του μηχανισμού, κάθε δείκτης στην ακολουθία μπορεί να υπολογίσει εξαρτήσεις εξετάζοντας ταυτόχρονα τόσο το παρελθόν όσο και το μέλλον. Η προσέγγιση αυτή προσφέρει στο μοντέλο μια πλήρη, σφαιρική εικόνα του περιβάλλοντος πλαισίου, βελτιστοποιώντας την εξαγωγή σύνθετων σημασιολογικών σχέσεων.

2) Αιτιώδεις Τοπολογίες (Causal Topologies)

Αντίθετα, κατά τη μοντελοποίηση γενετικών αυτοπαλίνδρομων ακολουθιών (generative autoregressive sequences), οι αρχιτεκτονικές οφείλουν να επιβάλλουν μια αυστηρή αιτιώδη μάσκα (causal mask) στη βάση του ιστορικού προτύπου που εισηγήθηκαν οι [Vaswani et al. \(2017\)](#). Στόχος της σχεδιαστικής αυτής επιλογής είναι ο εξαναγκασμός του δικτύου να σεβαστεί το βέλος του φυσικού χρόνου. Στο επίπεδο της αρχιτεκτονικής υλοποίησης, η μαθηματική γενίκευση της κλιμακωτής προσοχής εσωτερικού γινομένου με την ενσωμάτωση ενός κατευθυντικού πίνακα κάλυψης M (mask matrix) τυποποιείται στη σύγχρονη βιβλιογραφία ως:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V} \quad (19)$$

Στη μαθηματική αυτή διατύπωση, οι τιμές του πίνακα κάλυψης ορίζονται ως:

- $M_{ij} = 0$ για $i \geq j$
- $M_{ij} = -\infty$ για $i < j$

Η εισαγωγή της τιμής $-\infty$ μηδενίζει πρακτικά τις πιθανότητες της συνάρτησης softmax για όλα τα επόμενα στοιχεία. Αυτός ο μαθηματικός περιορισμός εγγυάται ότι η πρόβλεψη ενός ακουστικού διακριτικού (token) στο χρονικό βήμα t εξαρτάται αποκλειστικά από το ιστορικό πλαίσιο $u_{<t}$. Έτσι, διατηρείται η χρονολογική συνέπεια κατά τη διάρκεια της διαδοχικής αποκωδικοποίησης στον χρόνο εκτέλεσης (inference), εμποδίζοντας το μοντέλο να αποκτήσει πρόσβαση σε πληροφορίες από το μέλλον.

2.3.4 Ακουστική Προσαρμογή: Χαρτογράφηση Ακολουθίας με Επαυξημένη Συνέλιξη

Η αυτοπροσοχή πολλαπλών κεφαλών (Multi-Head Self-Attention) αναγνωρίζεται για την εξαιρετική της ικανότητα να μοντελοποιεί μακροσκοπικές και μακροδομικές εξαρτήσεις σε εκτεταμένες ακολουθίες. Παρά ταύτα, όπως επισημαίνουν οι [Gulati et al. \(2020\)](#), παρουσιάζει έναν σαφή αρχιτεκτονικό περιορισμό κατά την επεξεργασία ηχητικών

αναπαραστάσεων υψηλής πυκνότητας: την έλλειψη εντοπισμένης επαγωγικής πόλωσης (localized inductive bias). Τα αμιγή επίπεδα προσοχής αντιμετωπίζουν όλα τα στοιχεία της ακολουθίας ως ισαπέχοντα σημεία. Αυτή η ιδιότητα τα καθιστά αναποτελεσματικά στην απομόνωση γρήγορων ακουστικών μεταβατικών φαινομένων της κλίμακας των μικροδευτερολέπτων, των λεπτών αρμονικών του ύψους, καθώς και των τοπικών ευθυγραμμίσεων της φάσης του κύματος.

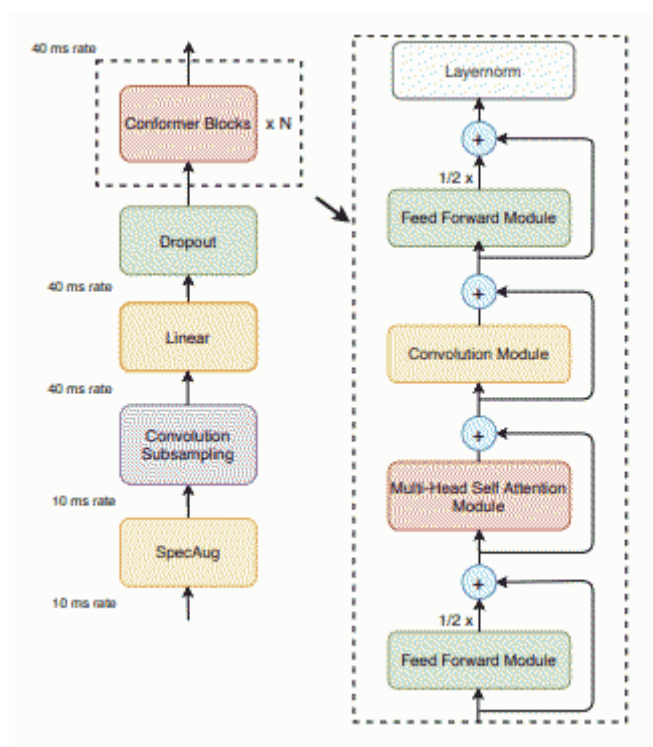
Για την υπέρβαση αυτών των περιορισμών στους ακουστικούς χώρους, οι σύγχρονοι διατροφικοί αγωγοί (pipelines) συχνά αντικαθιστούν τους τυπικούς Transformers με την αρχιτεκτονική Conformer, η οποία εισήχθη από τους [Gulati et al. \(2020\)](#). Η προσέγγιση αυτή αναδιαρθρώνει ριζικά το επίπεδο επεξεργασίας της ακολουθίας, οργανώνοντας τις υπολογιστικές λειτουργίες σε μια διάταξη δικτύου προώθησης (feed-forward network) τύπου Macaron, η οποία βασίζεται στο πρότυπο των [Lu et al. \(2019\)](#), παντρεύοντας την παγκόσμια προσοχή με τα τοπικά συνελκτικά επίπεδα, μέσα από μια ροή που απεικονίζεται αναλυτικά στην [Εικόνα 8](#).

Όπως παρατηρούμε στην εικόνα, η αρχιτεκτονική αναλύεται σε δύο επίπεδα:

- **Ο Αγωγός Εισόδου (Αριστερό Τμήμα):** Πριν το σήμα φτάσει στα κύρια υπολογιστικά μπλοκ, υφίσταται επαύξηση δεδομένων (SpecAug) και περνά από ένα επίπεδο υποδειγματοληψίας συνέλιξης (Convolution Subsampling). Το βήμα αυτό είναι κρίσιμο, καθώς συμπυκνώνει τη χρονική πυκνότητα του ήχου (από τα 10 ms στα 40 ms rate), καθιστώντας το σήμα διαχειρίσιμο.
- **Το Μπλοκ Conformer (Δεξί Τμήμα):** Αποτελεί την καρδιά του μοντέλου και υιοθετεί τη διάταξη τύπου **Macaron** (Lu et al., 2019). Αντί για ένα ενιαίο δίκτυο προώθησης, η υπολογιστική δομή «εγκλωβίζει» τις μονάδες προσοχής και συνέλιξης ανάμεσα σε δύο τροποποιημένα επίπεδα Feed-Forward, καθένα από τα οποία φέρει έναν συντελεστή κλιμάκωσης $1/2$ (half-step residual connections).

Συγκεκριμένα, η εσωτερική δομική ροή κάθε Conformer μπλοκ εκτελείται μέσω των ακόλουθων διαδοχικών λειτουργικών μονάδων:

- **Πρώτη Μονάδα Προώθησης (Feed-Forward Module - Ημί-Βήμα):** Προετοιμάζει και γραμμικοποιεί τις εισερχόμενες αναπαραστάσεις, εφαρμόζοντας το πρώτο μισό της δομής Macaron.
- **Μονάδα Αυτοπροσοχής Πολλαπλών Κεφαλών (Multi-Head Self-Attention):** Αναλαμβάνει την καταγραφή της καθολικής μακροδομής και των μακροπρόθεσμων εξαρτήσεων της ακολουθίας σε ολόκληρο το χρονικό εύρος.
- **Μονάδα Συνέλιξης (Convolution Module):** Επιστρατεύει διαχωρίσιμα κατά βάθος μονοδιάστατα συνελκτικά επίπεδα (1D CNNs) για να απομονώσει τα τοπικά ακουστικά μεταβατικά φαινόμενα, το ηχόχρωμα και τα χαρακτηριστικά υψηλής συχνότητας.
- **Δεύτερη Μονάδα Προώθησης (Feed-Forward Module - Ημί-Βήμα):** Ολοκληρώνει τη δομή Macaron, σταθεροποιώντας και εξάγοντας το τελικό διάνυσμα χαρακτηριστικών μέσω ενός τελικού επιπέδου κανονικοποίησης (Layernorm).



Εικόνα 8: Conformer encoder model architecture

Υπολογίζοντας παράλληλα την ολική αυτοπροσοχή και τις εντοπισμένες συνέλιξεις (οι οποίες παραμένουν αμετάβλητες στη μετατόπιση(shift-invariant), η αρχιτεκτονική καταγράφει ταυτόχρονα τόσο τις αφηρημένες γλωσσικές δομές όσο και τα τοπικά ηχητικά χαρακτηριστικά υψηλής πιστότητας. Ο συγκεκριμένος σχεδιασμός δημιουργεί μια εξαιρετικά στιβαρή ραχοκοκαλιά χαρτογράφησης ακολουθιών. Ελαχιστοποιεί αποτελεσματικά την απώλεια ή την παραμόρφωση πληροφορίας κατά τη μεταφορά των χαρακτηριστικών στους βαθύτερου επιπέδου αποκωδικοποιητές κυματομορφής (vocoders).

3. Ο αλγοριθμικός αγωγός για τη μετατροπή κειμένου σε ήχο

Η κατανόηση των τεχνικών μηχανικής μάθησης που διέπουν τη σύνθεση κειμένου σε ήχο (text-to-audio) προϋποθέτει τη ρήξη με μια παραδοσιακή αντίληψη: η παραγωγική διαδικασία δεν αποτελεί έναν μεμονωμένο, μονολιθικό υπολογισμό. Αντίθετα, όπως επισημαίνουν οι [Liu et al. \(2023\)](#), τα σύγχρονα συστήματα αιχμής λειτουργούν ως μια ενοποιημένη, πολυσταδιακή αλγοριθμική γραμμή επεξεργασίας (pipeline). Πρωταρχικός σκοπός αυτού του αγωγού επεξεργασίας είναι η λήψη μιας αραιής, εξαιρετικά αφηρημένης κειμενικής περιγραφής και η προοδευτική αποκωδικοποίησή της σε μια πυκνή, συνεχή ακουστική κυματομορφική αναπαράσταση πολλαπλών κλιμάκων, όπως παρουσιάζεται και στην [Εικόνα 9](#).

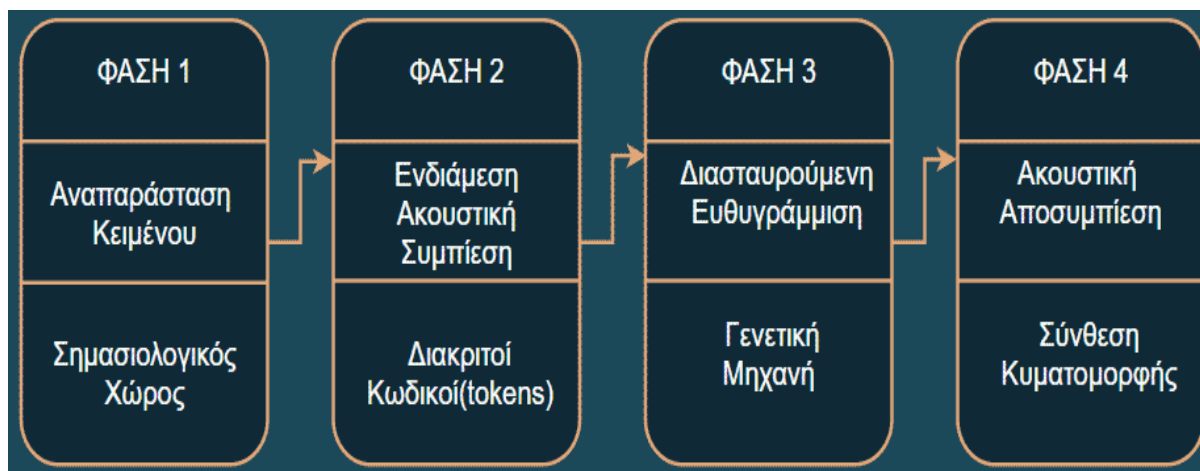
Πριν από τη διεξοδική μαθηματική και αρχιτεκτονική ανάλυση των επιμέρους στοιχείων, είναι απαραίτητη η χαρτογράφηση της καθολικής ροής των δεδομένων. Ο διατροφικός αυτός μετασχηματισμός, ο οποίος αναλύεται διεξοδικά στα έργα των [Liu et al. \(2023\)](#) και [Copet et al. \(2023\)](#), διαρθρώνεται σε τέσσερις θεμελιώδεις φάσεις:

- **Φάση 1: Σημασιολογική Εξαγωγή:** Η ακατέργαστη κειμενική προτροπή (text prompt) του χρήστη στερείται εγγενούς δομής, γεγονός που την καθιστά αρχικά μη αξιοποιήσιμη από μια μηχανή επεξεργασίας ήχου. Οι αλγόριθμοι του γλωσσικού front-end (linguistic front-end) απομονώνουν το κείμενο, το τεμαχίζουν σε επιμέρους γλωσσικές μονάδες και τις προβάλλουν σε έναν βαθύ χώρο ενσωμάτωσης συμφραζομένων (contextual embedding space), στη βάση του μεθοδολογικού προτύπου των [Peters et al. \(2018\)](#). Αυτό το στάδιο σταθεροποιεί την αφηρημένη ανθρώπινη πρόθεση, παράγοντας έναν πυκνό πίνακα σημασιολογικών διανυσμάτων, όπου οι πολιτισμικοί, συναισθηματικοί και υφολογικοί περιγραφείς μεταφράζονται σε σαφείς γεωμετρικές συντεταγμένες.
- **Φάση 2: Ακουστική Συμπύεση:** Παράλληλα, το σύστημα οφείλει να διαμορφώσει έναν χώρο-στόχο για τον ήχο, ο οποίος να αποτρέπει την υπολογιστική εξάντληση των πόρων. Λόγω του ότι ο ακατέργαστος ήχος εμπεριέχει δεκάδες χιλιάδες δείγματα

ανά δευτερόλεπτο, οι σύγχρονες αρχιτεκτονικές επιστρατεύουν νευρωνικούς κωδικοποιητές ήχου (neural audio codecs) για να συμπίεσουν την κυματομορφική είσοδο σε συμπαγείς λανθάνουσες αναπαράστασεις χαμηλής διάστασης, όπως καταδεικνύεται από τους [Défossez et al. \(2022\)](#). Η φάση αυτή εξαλείφει τις ακουστικές περιττότητες. Μετατρέπει εκτεταμένα ηχητικά σήματα σε μια αυστηρά οργανωμένη αλληλουχία διακριτών κωδικών (audio tokens) ή δισδιάστατων φασματικών μητρών. Έτσι, διασφαλίζεται ένας διαχειρίσιμος όγκος δεδομένων, τον οποίο το μοντέλο επεξεργάζεται ρεαλιστικά, χωρίς να προκαλείται υπερφόρτωση της μνήμης των GPU.

- **Φάση 3: Διασταυρούμενη Ευθυγράμμιση:** Αυτό το στάδιο αποτελεί τη λειτουργική καρδιά της αρχιτεκτονικής. Έχοντας απομονώσει τα σημασιολογικά διανύσματα κειμένου από τη μία πλευρά και τον συμπίεσμένο λανθάνοντα χώρο (latent space) του ήχου από την άλλη, εισάγεται η γενετική μηχανή —συνήθως ένας ισχυρός Transformer ή ένα μοντέλο λανθάνουσας διάχυσης (latent diffusion model), σύμφωνα με τις προσεγγίσεις που αναδεικνύουν οι προαναφερθέντες [Copet et al. \(2023\)](#). Μέσω επιπέδων διασταυρούμενης προσοχής (cross-attention layers), το μοντέλο εκπαιδεύεται στις στατιστικές συσχετίσεις ανάμεσα στους δύο ανόμοιους κόσμους, ακολουθώντας τη μαθηματική τυποποίηση των [Vaswani et al. \(2017\)](#). Τα διανύσματα κειμένου λειτουργούν ως καθοδηγητικός οδηγός: το σύστημα υπολογίζει δυναμικά τον τρόπο κατασκευής μιας νέας ακολουθίας ακουστικών χαρακτηριστικών ή αφαιρεί επαναληπτικά τον θόρυβο από έναν κενό καμβά, μέχρι να αναδυθεί μια συνεκτική ηχητική δομή.
- **Φάση 4: Ακουστική Αποκωδικοποίηση:** Στην τελική φάση, τα παραχθέντα χαρακτηριστικά βρίσκονται ακόμη εγκλωβισμένα σε έναν αφηρημένο, «σιωπηλό» λανθάνοντα χώρο. Οι αλγόριθμοι του back-end αναλαμβάνουν αυτές τις νεοσυντεθείσες αναπαραστάσεις και τις διοχετεύουν σε νευρωνικούς αποκωδικοποιητές και ψηφιακούς κωδικοποιητές φωνής (vocoders). Η διεργασία αυτή εκτελεί την αντίστροφη λειτουργία της δεύτερης φάσης: αποσυμπίεζει τις

μαθηματικές συντεταγμένες, αποκαθιστά τις λεπτές ακουστικές υφές, επιβάλλει τη σωστή ευθυγράμμιση της φάσης(phase alignment) και συνθέτει το τελικό, αναπαραγώγιμο ψηφιακό αρχείο ήχου υψηλής πιστότητας, βάσει των προτύπων σύνθεσης που εισηγήθηκαν οι [Kong et al. \(2020\)](#).



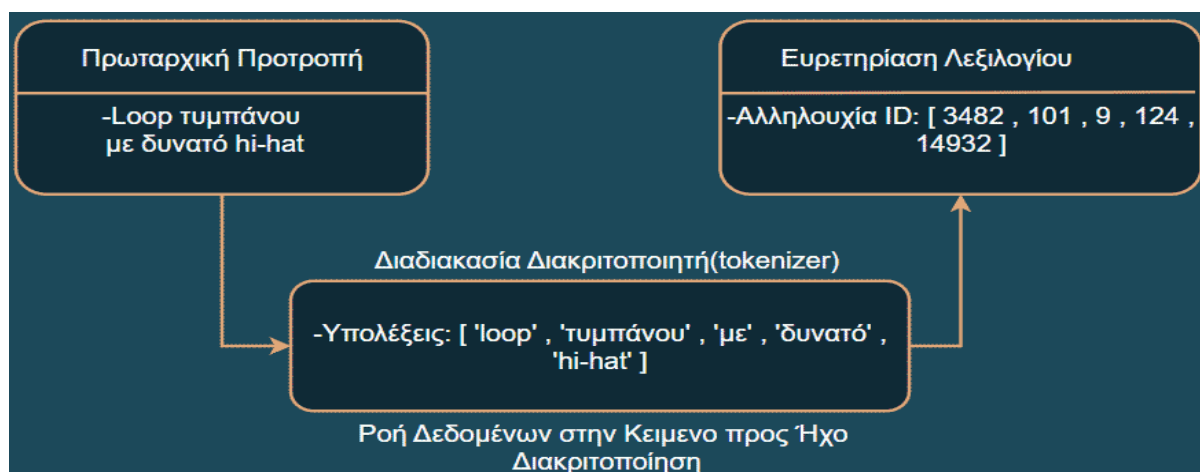
Εικόνα 9: Σχηματική περιγραφή του αλγοριθμικού αγωγού απο κείμενο σε ήχο

3.1 Εξαγωγή Χαρακτηριστικών Front-End (Αλγόριθμοι Κειμενοποίησης)

Η λειτουργική εκτέλεση της μετατροπής κειμένου σε ήχο ξεκινά στη διεπαφή μεταξύ της ανθρώπινης φυσικής γλώσσας και του περιβάλλοντος των νευρωνικών δικτύων που βασίζεται σε tensor(πολυδιάστατοι πίνακες δεδομένων). Ο σκοπός αυτής της ενότητας είναι να μετατρέψει ακατέργαστες, αδόμητες προτροπές κειμενικής προετοιμασίας - όπως μουσικές περιγραφές, περιορισμούς είδους ή παραμέτρους ακουστικού περιβάλλοντος - σε πυκνούς tensor σημασιολογικής προετοιμασίας. Αυτή η διαδικασία απαιτεί μια εξαιρετικά βελτιστοποιημένη ακολουθία διακριτοποίησης(tokenization), χαρτογράφηση ευρετηρίου λεξικού και προβολή υψηλής διάστασης μέσω "παγωμένων" δικτύων κωδικοποιητών κειμένου.

3.1.1 Αλγοριθμική Τεμαχιοποίηση και Χαρτογράφηση Λεξιλογίου

Πριν από τον υπολογισμό των σημασιολογικών αλληλεπιδράσεων, η προτροπή κειμένου (text prompt) διέρχεται από έναν ντετερμινιστικό διαχωριστή υπολέξεων (subword tokenizer). Οι σύγχρονες αρχιτεκτονικές μετατροπής κειμένου σε ήχο (text-to-audio pipelines) βασίζονται σε μοντέλα τεμαχιοποίησης που προέρχονται από γλωσσικές αρχιτεκτονικές μεγάλης κλίμακας. Κεντρικό παράδειγμα αποτελεί το ανοιχτό πλαίσιο SentencePiece, το οποίο αναπτύχθηκε από τον [Kudo \(2018\)](#) και λειτουργεί ανεξάρτητα από τη γλώσσα εισαγωγής.



Εικόνα 10: Μετατροπή προτροπής σε tokens.

Η διαδικασία αυτή, όπως φαίνεται εικονοποιημένα στην [Εικόνα 10](#), μετατρέπει την αυθαίρετη συμβολοσειρά εισόδου σε μια διακριτή ακολουθία ακέραιων αναγνωριστικών (token IDs) $t = [t_1, t_2, \dots, t_L]$. Στη μαθηματική αυτή διατύπωση, η μεταβλητή L ορίζει το μήκος της ακολουθίας. Κάθε στοιχείο $t_i \in \{1, 2, \dots, |V|\}$ αντιστοιχεί σε μια συγκεκριμένη θέση στο προκαθορισμένο προϋπολογισμό του μεγέθους λεξιλογίου $|V|$.

Λειτουργώντας σε επίπεδο υπολέξεων ή χαρακτήρων-byte, ο μηχανισμός εξαλείφει πλήρως το σημείο συμφόρησης που προκαλούν τα σφάλματα εκτός λεξιλογίου (Out-of-Vocabulary). Σύνθετοι ακουστικοί ή μουσικολογικοί όροι αναλύονται συστηματικά σε αναγνωρίσιμα μορφηματικά κομμάτια. Αυτό διασφαλίζει τη σταθερή ροή της πληροφορίας προς τα επόμενα νευρωνικά επίπεδα.

3.1.2 Αρχιτεκτονικά Παραδείγματα Κωδικοποίησης Κειμένου

Μετά την παραγωγή της ακολουθίας των διακριτών αναγνωριστικών t (token Ids), η πληροφορία διοχετεύεται σε έναν κωδικοποιητή κειμένου (text encoder) που βασίζεται σε βαθιές αρχιτεκτονικές Transformer, ακολουθώντας το γενικό πρότυπο των [Raffel et al. \(2020\)](#). Στο πεδίο των σύγχρονων συστημάτων μετατροπής κειμένου σε ήχο (text-to-audio) και κειμένου σε μουσική (text-to-music), δύο διακριτές φιλοσοφίες διαμορφώνουν τη διατροφική εξαρτημένη επεξεργασία (conditioning):

1) Πυκνοί Σημασιολογικοί Κωδικοποιητές (T5 / LLMs)

Όταν η έμφαση δίνεται στη λεπτομερή σύνταξη και τον ακριβή έλεγχο σύνθετων οδηγιών, η χρήση καθαρών γλωσσικών κωδικοποιητών, όπως ο Text-to-Text Transfer Transformer (T5), κρίνεται ιδανική. Το μοντέλο T5 αξιοποιεί έναν αμφίδρομο κωδικοποιητή (bidirectional encoder), προεκπαιδευμένο σε τεράστια σώματα κειμένων, γεγονός που του επιτρέπει να αναγνωρίζει λεπτές γραμματικές ιεραρχίες και αφηρημένες πολιτισμικές έννοιες. Κατά την ενσωμάτωσή του σε γενετικούς αγωγούς όπως το MusicGen των [Copet et al. \(2023\)](#), ο κωδικοποιητής επεξεργάζεται την ακολουθία t μέσω διαδοχικών επιπέδων αυτοπροσοχής. Στο επίπεδο της αρχιτεκτονικής υλοποίησης, η εξαγωγή του συνεχούς πίνακα κρυφών καταστάσεων (hidden states) μοντελοποιείται συμβολικά στη σύγχρονη βιβλιογραφία ως:

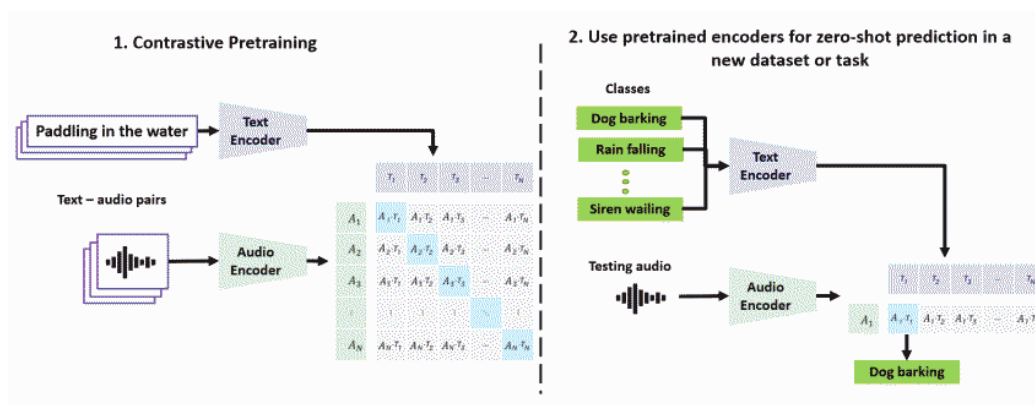
$$\mathbf{H}_{text} = f_{T5}(t) \in \mathbb{R}^{L \times D_{text}} \quad (20)$$

Η μεταβλητή D_{text} ορίζει την κρυφή διάσταση του κωδικοποιητή, η οποία συνήθως λαμβάνει τιμές όπως 1024 ή 4096. Ο πίνακας αυτός εξαρτάται άμεσα από το μήκος της ακολουθίας, διατηρώντας ανέπαφη τη χωρική και διαδοχική ευθυγράμμιση των επιθέτων με τα αντίστοιχα ουσιαστικά. Ο μηχανισμός διασφαλίζει μαθηματικά ότι ο προσδιορισμός «loop» παραμένει συνδεδεμένος με τη λέξη «τυμπάνου» και όχι με τη λέξη «δυνατό».

2) Αντιπαραθετικοί Κωδικοποιητές Κοινού Χώρου Ήχου-Κειμένου (CLAP)

Μια εναλλακτική προσέγγιση παρακάμπτει τη λεπτομερή γλωσσική σύνταξη για να δώσει προτεραιότητα στην άμεση διατροπική ευθυγράμμιση (cross-modal alignment). Αυτό επιτυγχάνεται μέσω αρχιτεκτονικών Αντιπαραθετικής Προεκπαίδευσης Γλώσσας-Ήχου (Contrastive Language-Audio Pretraining - CLAP), όπως αυτές αναλύονται στα έργα των [Elizalde et al. \(2023\)](#) και [Wu et al. \(2023\)](#).

Η δομή τους ακολουθεί τη φιλοσοφία των μοντέλων CLIP του οπτικού τομέα που εισηγήθηκαν οι [Radford et al. \(2021\)](#). Ο κωδικοποιητής κειμένου CLAP εκπαιδεύεται μέσω ενός συστήματος διπλού κωδικοποιητή (dual-encoder system) παράλληλα με έναν κωδικοποιητή ήχου (audio encoder), μια διπλή υπολογιστική ροή που αποτυπώνεται στην [Εικόνα 11](#).



Εικόνα 11: Αρχιτεκτονική και διαδικασία λειτουργίας του μοντέλου CLAP

Σε αυτή την περίπτωση, αντί για έναν πλήρη πίνακα αναπαραστάσεων, ο κωδικοποιητής συμπυκνώνει την ακολουθία των διακριτικών σε ένα ενιαίο, καθολικά συγκεντρωμένο διάνυσμα ενσωμάτωσης (globally pooled embedding vector). Στην τρέχουσα βιβλιογραφία, η μαθηματική γενίκευση αυτής της διαδικασίας συμπύκνωσης τυποποιείται ως:

$$\mathbf{z}_{text} = f_{CLAP}(\mathbf{t}) \in \mathbb{R}^{D_{joint}} \quad (21)$$

Αυτή η προβολή συμπεριέχει ολόκληρη την προτροπή σε μια μοναδική σημασιολογική συντεταγμένη. Όπως παρουσιάζεται στην [Εικόνα 11](#), η αρχιτεκτονική αυτή λειτουργεί και αξιοποιείται σε δύο διακριτές φάσεις:

- **1. Αντιπαραθετική Προεκπαίδευση (Contrastive Pretraining):** Κατά την εκπαίδευση (αριστερό τμήμα), ζεύγη κειμένου και ήχου τροφοδοτούνται παράλληλα στους αντίστοιχους κωδικοποιητές. Οι αναπαραστάσεις κειμένου ($T_1, \dots, T_N$ που αντιστοιχούν στο διάνυσμα z_{text}) και οι αναπαραστάσεις ήχου ($A_1, \dots, A_N$) σχηματίζουν έναν πίνακα ομοιότητας. Χρησιμοποιώντας την αντιπαραθετική συνάρτηση απώλειας InfoNCE, το σύστημα εξαναγκάζει τα σωστά ζευγάρια (η κύρια διαγώνιος $A_i T_i$) να αποκτήσουν μέγιστη ομοιότητα συνημίτονου (cosine similarity), ενώ απωθεί όλες τις άλλες άσχετες συνδυαστικές κυψέλες.
- **2. Πρόβλεψη Μηδενικού Δείγματος (Zero-Shot Prediction):** Μετά την προεκπαίδευση (δεξί τμήμα), οι «παγωμένοι» πλέον κωδικοποιητές μπορούν να χρησιμοποιηθούν άμεσα για ταξινόμηση ή ποιοτική αξιολόγηση χωρίς επιπλέον εκπαίδευση. Ένα άγνωστο δείγμα ήχου (Testing audio) κωδικοποιείται σε ένα διάνυσμα A_I και συγκρίνεται ταυτόχρονα με τα διανύσματα κειμένου που προκύπτουν από διάφορες υποψήφιες ετικέτες/κλάσεις (π.χ. "Dog barking", "Rain falling"). Η κλάση με την υψηλότερη βαθμολογία ομοιότητας επιλέγεται ως η τελική πρόβλεψη.

Σε αγωγούς λανθάνουσας διάχυσης (latent diffusion pipelines), όπως το AudioLDM των [Liu et al. \(2023\)](#), η καθολική αυτή ενσωμάτωση z_{text} λειτουργεί ως οδηγός υψηλού επιπέδου για το δίκτυο αποθορυβοποίησης (denoising network). Καθορίζει με ακρίβεια το γενικό ύφος, την ακουστική ατμόσφαιρα και το θεματικό περιεχόμενο της παραγόμενης ακουστικής σύνθεσης.

3.1.3 Η στρατηγική των παγωμένων βαρών και της εξαγωγής χαρακτηριστικών

Όπως επισημαίνουν οι [Liu et al. \(2023\)](#), το πλήρες «πάγωμα» των παραμέτρων του κωδικοποιητή κειμένου θ_{text} αποτελεί θεμελιώδη μηχανική επιλογή στα σύγχρονα συστήματα αιχμής. Κατά την εκπαίδευση του παραγωγικού μοντέλου, ο κωδικοποιητής του front-end παραμένει αυστηρά στατικός. Οι διαβαθμίσεις της οπισθοδιάδοσης (backpropagation gradients) υπολογίζονται και εφαρμόζονται αποκλειστικά στα βαθύτερα επίπεδα σύνθεσης του ήχου, σταματώντας ουσιαστικά τη ροή των ανανεώσεων στην ακουστική μηχανή.

Η αρχιτεκτονική αυτή προσέγγιση εξασφαλίζει τρία καθοριστικά πλεονεκτήματα:

- **Μετριασμός της μετατόπισης αναπαράστασης (representation drift):** Η σταθερότητα των βαρών εμποδίζει το γλωσσικό μοντέλο από το να αλλοιώσει τη δομή του. Χωρίς αυτή τη στρατηγική, οι εσωτερικοί γλωσσικοί κανόνες θα υπερπροσαρμόζονταν (overfit) στο σαφώς μικρότερο σύνολο δεδομένων ήχου-κειμένου. Το πάγωμα διασφαλίζει ότι η ευρεία, προεκπαιδευμένη σημασιολογική γνώση του μοντέλου για τον κόσμο παραμένει ανέπαφη ([Liu et al. \(2023\)](#)).
- **Αποτροπή της καταστροφικής λήθης (catastrophic forgetting):** Διασφαλίζεται η διατήρηση των δυνατοτήτων γενίκευσης μηδενικής διδασκαλίας (zero-shot generalization). Η αρχιτεκτονική διατηρεί την ικανότητα να ερμηνεύει σπάνιες, σύνθετες ή εξαιρετικά αφηρημένες κειμενικές προτροπές (text prompts). Το σύστημα ανταποκρίνεται με επιτυχία ακόμη και σε περιγραφές που δεν υπήρχαν ποτέ ρητά στα περιορισμένα δεδομένα εκπαίδευσης του ήχου.
- **Δραστική μείωση του υπολογιστικού κόστους:** Στη βάση των προσεγγίσεων που εισήγαγαν οι [Raffel et al. \(2020\)](#) και [Elizalde et al. \(2023\)](#), οι αναπαραστάσεις κειμένου (H_{text} ή z_{text}) υπολογίζονται εκ των προτέρων (pre-computation) μία μόνο φορά ανά παρτίδα (batch).

Οι εξαχθείσες αναπαραστάσεις αποθηκεύονται στη συνέχεια στην προσωρινή μνήμη (caching). Με τον τρόπο αυτό, το δίκτυο αποφεύγει τα εξαιρετικά κοστοβόρα περάσματα

οπισθοδιάδοσης μέσα από δισεκατομμύρια παραμέτρους του γλωσσικού κωδικοποιητή. Η πολύτιμη μνήμη των GPU απελευθερώνεται και διατίθεται αποκλειστικά για τις απαιτητικές εργασίες σύνθεσης ήχου υψηλής διαστατικότητας.

3.1.4 Αγωγοί Επαύξησης Κειμένου και Παραγωγής Λεζαντών (Captioning)

Η μεταβλητή ποιότητα και η σπανιότητα των κειμενικών περιγραφών στα διαδικτυακά σύνολα δεδομένων (datasets) αποτελεί σημαντική λειτουργική πρόκληση για την εκπαίδευση των μοντέλων μετατροπής κειμένου σε ήχο. Αντί για ρέουσες περιγραφικές προτάσεις, τα σώματα εκπαίδευσης εμπεριέχουν συχνά ελλιπή μεταδεδομένα (metadata) και αποσπασματικές ετικέτες (tags), περιορίζοντας δραματικά τη διαθέσιμη πληροφορία. Όπως καταδεικνύεται από τους [Wu et al. \(2023\)](#), οι σύγχρονοι αγωγοί του front-end επιλύουν αυτό το δομικό πρόβλημα εισάγοντας ένα αυτοματοποιημένο στάδιο γλωσσικής επαύξησης (text augmentation). Συγκεκριμένα, πριν από τη διαδικασία του τεμαχισμού σε υπολέξεις (subword tokenization), επιστρατεύονται στατικά, «παγωμένα» Μεγάλα Γλωσσικά Μοντέλα (Large Language Models - LLMs), ακολουθώντας τη μεθοδολογική προσέγγιση που περιγράφουν οι [Liu et al. \(2023\)](#).

Κατά την προεπεξεργασία, το σύστημα τροφοδοτεί τις φτωχές ετικέτες σε ένα LLM, το οποίο καθοδηγείται από εξειδικευμένα πρότυπα προτροπών (prompt templates). Το γλωσσικό μοντέλο μετατρέπει τις απλές λέξεις-κλειδιά σε ολοκληρωμένες, δομημένες λεζάντες (captions). Οι νέες αυτές περιγραφές αποτυπώνουν με ακρίβεια ακουστικά ηχοχρώματα, χωρικές ιδιότητες του περιβάλλοντος, ενορχηστρώσεις και συναισθηματικές αποχρώσεις. Η συνθετική αυτή επέκταση διασφαλίζει ότι ο κωδικοποιητής κειμένου θα λάβει ένα εξαιρετικά πυκνό σημασιολογικό πεδίο. Έτσι, τα βαθύτερα επίπεδα διασταυρούμενης προσοχής (cross-attention) αποκτούν πληθώρα συμφραζομένων αναφοράς. Αυτό επιτρέπει στο δίκτυο να αντιστοιχίζει πλέον με απόλυτη ακρίβεια τα γλωσσικά χαρακτηριστικά στα κατάλληλα ακουστικά και οργανικά στοιχεία του λανθάνοντος χώρου .

3.1.5 Μηχανική Άνευ Όρων Προετοιμασίας και Καθοδήγηση Χωρίς Ταξινομητές (CFG)

Όπως τεκμηριώνεται από τον [Ho \(2022\)](#), η λειτουργία του γλωσσικού front-end καθορίζει άμεσα την ελεγχσιμότητα του παραγωγικού μοντέλου μέσω της τεχνικής της Καθοδήγησης Χωρίς Ταξινομητές (Classifier-Free Guidance - CFG). Προκειμένου το σύστημα να ελέγχει δυναμικά τον βαθμό συμμόρφωσης με την κειμενική προτροπή κατά τον χρόνο εκτέλεσης (inference), η εξαγωγή χαρακτηριστικών δεν πρέπει να λειτουργεί ως μια απλή, ντετερμινιστική χαρτογράφηση. Το μοντέλο οφείλει να αναπαριστά ενεργά και την άνευ όρων (unconditioned) κατάσταση κατά τη φάση της εκπαίδευσης.

Κατά την εκπαίδευση, ο αγωγός εφαρμόζει έναν μηχανισμό τυχαίας απόρριψης (stochastic dropout) στους ταυστές (tensors) των κειμενικών χαρακτηριστικών (H_{text} ή z_{text}). Η διαδικασία αυτή ακολουθεί μια σταθερή πιθανότητα p_{drop} , η οποία συνήθως κυμαίνεται μεταξύ 0,10 και 0,20. Στο επίπεδο της υλοποίησης της αρχιτεκτονικής, η στοχαστική αυτή αντικατάσταση μορφοποιείται μαθηματικά στη σύγχρονη βιβλιογραφία μέσω της ακόλουθης συμβολικής μοντελοποίησης:

$$H_{conditioned} = \begin{cases} H_{text} & \text{with probability } 1 - p_{drop} \\ H_{\emptyset} & \text{with probability } p_{drop} \end{cases} \quad (22)$$

Στην παραπάνω διανυσματική διατύπωση, η μεταβλητή $H_{\emptyset}$ αντιπροσωπεύει μια εκπαιδύσιμη, στατική ενσωμάτωση κενού κειμένου (null-prompt embedding), η αναπαράσταση αυτή διατηρεί τις ίδιες ακριβώς διαστάσεις με τις πραγματικές ενσωματώσεις κειμένου.

Με αυτή τη σκόπιμη παράλειψη της κειμενικής πληροφορίας, ο γενετικός πυρήνας εξαναγκάζεται να μάθει ταυτόχρονα δύο διαφορετικές κατανομές: την υπό όρους (conditional) χαρτογράφηση κειμένου-ήχου και την άνευ όρων (unconditional) κατανομή του ακουστικού σήματος. Στη φάση του inference, η μηχανή υπολογίζει μια γραμμική παρέκταση (linear extrapolation) ανάμεσα στις δύο αυτές καταστάσεις. Αυξάνοντας το βάρος της

διαφοράς τους μέσω ενός συντελεστή κλίμακας (guidance scale), το σύστημα ενισχύει επιλεκτικά τα ακουστικά χαρακτηριστικά που ευθυγραμμίζονται με το κείμενο, ενώ ταυτόχρονα καταστέλλει τον γενικό ακουστικό θόρυβο.

3.2 Διακριτοποίηση και Ακουστική Συμπύεση (Νευρωνικοί Κωδικοποιητές Ήχου)

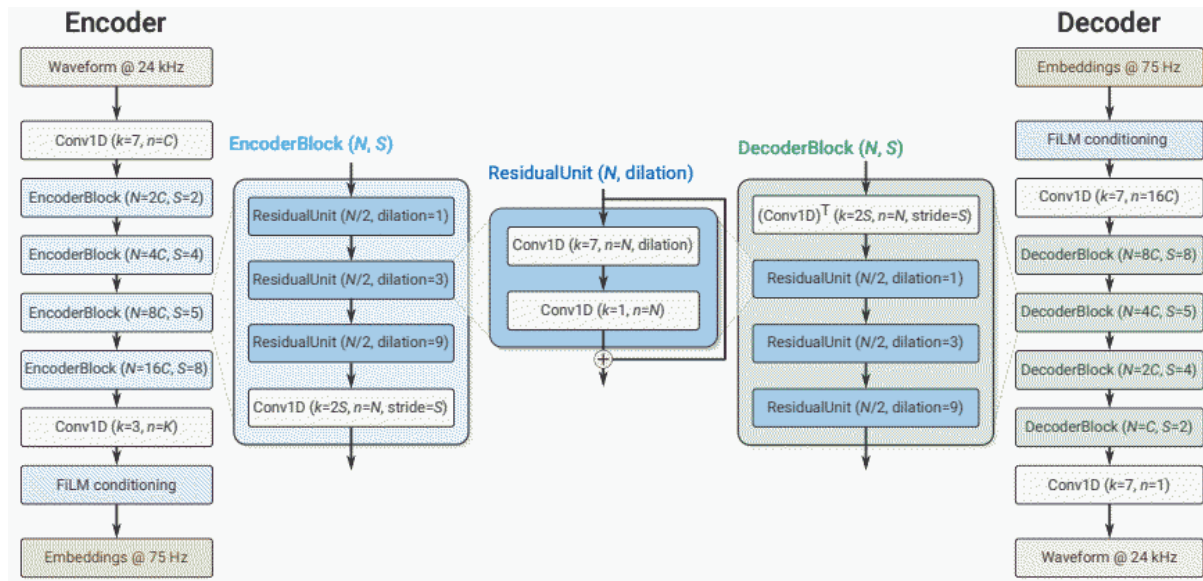
Για να γεφυρωθεί το τεράστιο χάσμα διαστάσεων μεταξύ των συμβολοσειρών κειμένου και των ακατέργαστων κυματομορφών, τα σύγχρονα συστήματα μετατροπής κειμένου σε ήχο απαιτούν ένα δυναμικό σημείο συμφόρησης στη συμπύεση δεδομένων. Αυτό το σημείο συμφόρησης δημιουργείται μέσω νευρωνικών κωδικοποιητών ήχου. Λειτουργώντας ως βαθιά συνελκτικοί αυτόματοι κωδικοποιητές, αυτά τα δίκτυα απορροφούν συνεχείς κυματομορφές ήχου, τις προβάλλουν μέσω ενός λανθάνοντος καναλιού χαμηλής συχνότητας, τις διακριτοποιούν σε ακέραιες ακολουθίες και τις αποκωδικοποιούν ξανά σε ακατέργαστα ηχητικά κύματα με ελάχιστη απώλεια αντίληψης. Στον πυρήνα αυτού του μηχανισμού συμπύεσης βρίσκονται οι αλγόριθμοι Κβαντοποίησης Διανύσματος (VQ) και Κβαντοποίησης Υπολειμματικού Διανύσματος (RVQ), αλλά και σύγχρονοι όπως η Πεπερασμένη Βαθμωτή Κβαντοποίηση (FSQ) και η Κανονικοποίηση Υπερσφαιρών (Hypersphere Normalization).

3.2.1 Η θεμελιώδης αρχιτεκτονική των νευρωνικών κωδικοποιητών ήχου

Όπως τεκμηριώνεται στα έργα των [Zeghidour et al. \(2021\)](#) και [Défossez et al. \(2022\)](#), η ακατέργαστη κυματομορφή ήχου $x \in \mathbb{R}^{T \times C}$ παρουσιάζει εξαιρετικά υψηλό βαθμό πλεονασμού (όπου T είναι το πλήθος των χρονικών δειγμάτων και C ο αριθμός των ακουστικών καναλιών). Την απαραίτητη μείωση της ακραίας αυτής διαστατικότητας αναλαμβάνει ο κωδικοποιητής f_{enc} του νευρωνικού δικτύου. Η μικρο-αρχιτεκτονική αυτού του συστήματος βασίζεται σε διαδοχικά μονοδιάστατα συνελκτικά επίπεδα υποδειγματοληψίας (1D downsampling convolutions) τα οποία συνδυάζονται με διευρυμένα υπολειμματικά μπλοκ (dilated residual blocks), όπως απεικονίζεται λεπτομερώς στο δομικό Διπλωματική Εργασία

διάγραμμα της [Εικόνας 12](#). Η διαδικασία αυτή μπορεί να αποδοθεί συμβολικά ως η χαρτογράφηση του ακατέργαστου σήματος σε μια συνεχή λανθάνουσα αναπαράσταση:

$$\mathbf{z} = f_{enc}(\mathbf{x}) \in \mathbb{R}^{S \times D} \quad (23)$$



Εικόνα 12: Αρχιτεκτονική Κωδικοποιητή-Αποκωδικοποιητή

Η επεξεργασία του σήματος εκτελείται μέσω εξειδικευμένων υποενοτήτων:

- **EncoderBlock (N, S):** Κάθε κεντρικό μπλοκ του κωδικοποιητή δέχεται έναν αριθμό καναλιών N και εφαρμόζει ένα βήμα συνέλιξης (stride S). Στο εσωτερικό του, η πληροφορία διακλαδίζεται σε τρεις διαδοχικές υπομονάδες **ResidualUnit**, οι οποίες χρησιμοποιούν διευρυμένες συνελίξεις (dilated convolutions) με εκθετικά αυξανόμενο ρυθμό διεύρυνσης (dilation = 1, 3, 9). Η στρατηγική αυτή επιτρέπει στο δίκτυο να διευρύνει δραστικά το δεκτικό του πεδίο (receptive field) χωρίς να χάνει την τοπική χρονική ανάλυση.
- **Μηχανισμός FiLM (Feature-wise Linear Modulation):** Στο τέλος της υπολογιστικής αλυσίδας του κωδικοποιητή και στην αρχή του αποκωδικοποιητή,

εισάγεται ένα επίπεδο FiLM conditioning. Το επίπεδο αυτό εκτελεί μια γραμμική μεταβολή (scaling και shifting) στους χάρτες χαρακτηριστικών του δικτύου, επιτρέποντας την εισαγωγή εξωτερικών πληροφοριών ελέγχου στο μοντέλο.

- **Decoder:** Ο αποκωδικοποιητής ακολουθεί την ακριβώς αντίστροφη υπολογιστική πορεία, αντικαθιστώντας τις συνελίξεις υποδειγματοληψίας με επίπεδα αναστραμμένης συνέλιξης $[(ConvID)^T$ - transposed convolutions] για την ανακατασκευή του σήματος.

Καθοριστικό ρόλο στη διαδικασία παίζει το βήμα (stride) αυτών των συνελίξεων, καθώς συμπιέζει δραστικά τη χρονική διάσταση των πλαισίων (frames) επιτυγχάνοντας μια σχέση $S \ll T$. Όπως αποτυπώνεται στο συγκεκριμένο παράδειγμα του SoundStream, η αρχιτεκτονική μετατρέπει τον ακατέργαστο φυσικό ρυθμό μιας κυματομορφής των 24 kHz σε ένα εξαιρετικά αραιό και διαχειρίσιμο πλέγμα συνεχών ενσωματώσεων (embeddings) με ρυθμό μόλις 75Hz.

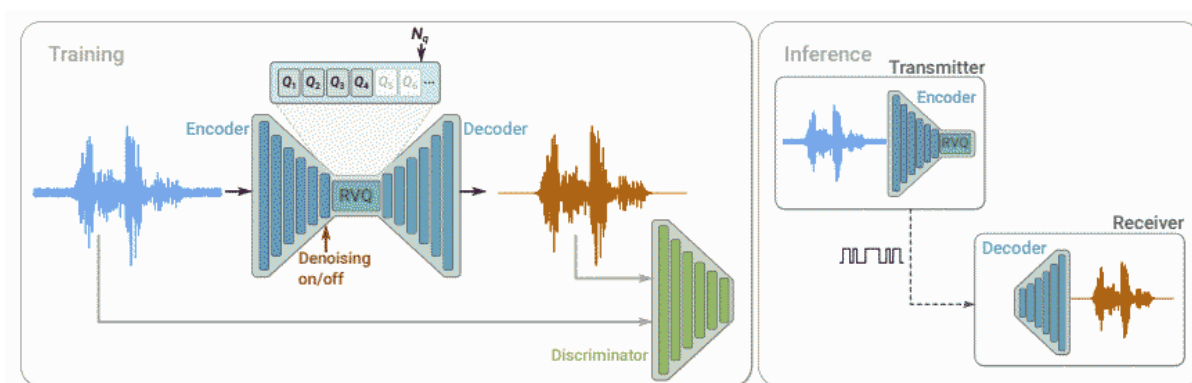
Παρά τη δραστική αυτή μείωση, ανακύπτει ένα νέο αρχιτεκτονικό εμπόδιο. Το διάνυσμα σε κάθε βήμα $s \in S$ εξακολουθεί να αποτελείται από συνεχείς μεταβλητές κινητής υποδιαστολής $z_s \in \mathbb{R}^D$. Οι αναπαραστάσεις αυτές είναι αδύνατον να επεξεργαστούν άμεσα από μοντέλα διακριτής μοντελοποίησης ακολουθιών, όπως οι Transformers. Για τη γεφύρωση αυτού του χάσματος, καθίσταται πλέον υποχρεωτική η εισαγωγή ενός επιπέδου κβαντισμού (quantization), η θεωρητική βάση του οποίου εδράζεται στο κλασικό πλαίσιο των van den Oord et al. (2017).

Η συνολική όμως ροή αυτής της αρχιτεκτονικής κωδικοποίησης, κβαντισμού και ανακατασκευής οπτικοποιείται στην [Εικόνα 13](#), η οποία διαχωρίζεται σε δύο λειτουργικές φάσεις:

- **1. Φάση Εκπαίδευσης (Training):** Κατά την εκπαίδευση (αριστερό τμήμα), η συνεχής λανθάνουσα αναπαράσταση περνά μέσα από ένα επίπεδο Υπολειμματικού Κβαντισμού Διανυσμάτων (Residual Vector Quantization - RVQ). Το RVQ

επιστρατεύει μια σειρά από n_q διακριτούς κβαντιστές ($Q_1, Q_2, \dots$), όπου κάθε επόμενος κβαντιστής μοντελοποιεί αποκλειστικά το υπόλοιπο (residual) σφάλμα του προηγούμενου. Το σήμα ανακατασκευάζεται από τον Decoder και οδηγείται, παράλληλα με τον αρχικό ήχο, σε ένα υποστηρικτικό δίκτυο Εξομοιωτή (Discriminator). Μέσω αντιπαραθετικών συναρτήσεων απώλειας (adversarial losses), ο Discriminator αναγκάζει τον Decoder να παράγει ήχο υψηλής φυσικότητας, χωρίς μεταλλικά τεχνουργήματα.

- **Σύνδεση με το Denoising:** Στη φάση αυτή, η είσοδος "Denoising on/off" καθοδηγεί το επίπεδο FiLM conditioning, εκπαιδεύοντας το μοντέλο να λειτουργεί δυναμικά τόσο ως απλός κωδικοποιητής όσο και ως φίλτρο αφαίρεσης θορύβου περιβάλλοντος.
- **2. Φάση Παραγωγής / Εφαρμογής (Inference):** Μετά την ολοκλήρωση της εκπαίδευσης (δεξί τμήμα), ο Discriminator αφαιρείται. Ο κωδικοποιητής και ο κβαντιστής RVQ που είναι εγκατεστημένοι στον πομπό (Transmitter) μετατρέπουν τον ήχο σε μια εξαιρετικά συμπίεσμένη διακριτή ροή bit (bitstream). Αυτή η ψηφιακή ακολουθία μπορεί να μεταδοθεί με ελάχιστες απαιτήσεις εύρους ζώνης και να ανακατασκευαστεί ακαριαία σε κυματομορφή από τον αποκωδικοποιητή του δέκτη (Receiver).



Εικόνα 13: Αρχιτεκτονική του νευρωνικού μοντέλου SoundStream

3.2.2 Μηχανική Κβαντοποίησης Διανυσμάτων (VQ)

Η διακριτή νευρωνική χαρτογράφηση του ακουστικού σήματος στηρίζεται κατά κύριο λόγο στον αλγόριθμο Κβαντοποίησης Διανυσμάτων (Vector Quantization). Η τεχνική αυτή προέρχεται από το αρχιτεκτονικό πλαίσιο του Vector Quantized Variational Autoencoder (VQ-VAE), το οποίο εισηγήθηκαν οι [van den Oord et al. \(2017\)](#). Το επίπεδο VQ εισάγει ένα διακριτό βιβλίο κωδικών (codebook) $C = \{e_1, e_2, \dots, e_K\}$ όπου η μεταβλητή K ορίζει το μέγεθος του λεξιλογίου (συνήθως 1024 ή 2048). Κάθε καταχώρηση $e_i \in \mathbb{R}^D$ αποτελεί ένα εκπαιδεύσιμο, συνεχές διάνυσμα ενσωμάτωσης.

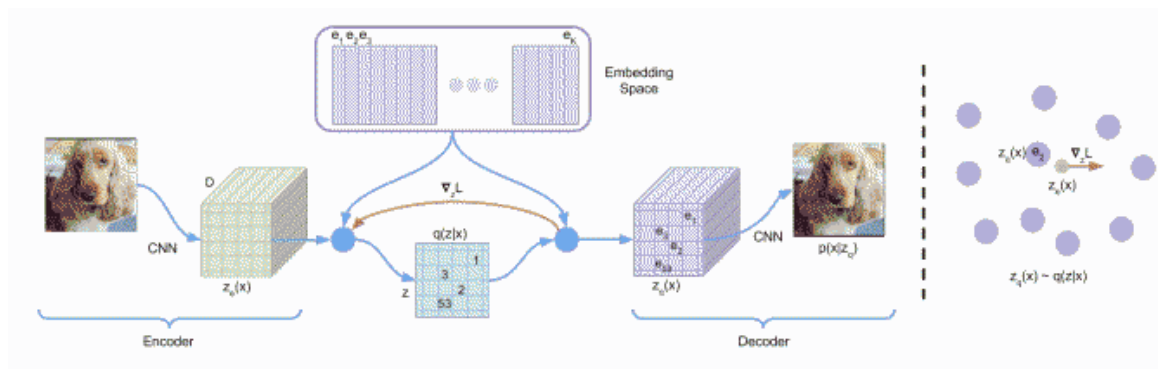
Για κάθε συνεχές λανθάνον διάνυσμα z_s που παράγει ο κωδικοποιητής, ο τελεστής κβαντοποίησης $q(\cdot)$ εκτελεί μια ντετερμινιστική αναζήτηση πλησιέστερου γείτονα (nearest neighbor search) σε ολόκληρο το codebook. Ο αλγόριθμος υπολογίζει την Ευκλείδεια απόσταση ανάμεσα στο διάνυσμα z_s και σε όλες τις καταχωρήσεις του βιβλίου, απομονώνοντας τον δείκτη της πλησιέστερης συντεταγμένης. Η διαδικασία αυτή τυποποιείται στη σύγχρονη βιβλιογραφία ως:

$$z_{q,s} = q(z_s) = e_k \quad \text{όπου} \quad k = \underset{i}{\operatorname{argmin}} \|z_s - e_i\|_2 \quad (24)$$

Όπως απεικονίζεται σχηματικά στο γενικό αρχιτεκτονικό διάγραμμα της [Εικόνας 14](#) (αριστερό τμήμα), η ροή εκτελείται ως εξής:

- **Κωδικοποίηση:** Το σήμα εισόδου περνά από έναν κωδικοποιητή (Encoder) που παράγει τον συνεχή πίνακα αναπαραστάσεων (στο διάγραμμα σημειώνεται ως $z_e(x)$), ο οποίος ταυτίζεται με τη μεταβλητή z_s του κειμένου).
- **Κβαντισμός:** Ο τελεστής $q(z|x)$ χαρτογραφεί κάθε διάνυσμα στην πλησιέστερη καταχώρηση του Embedding Space. Η έξοδος αυτής της διαδικασίας είναι διπλή: παράγεται μια ακολουθία διακριτών ακεραίων δεικτών (z), η οποία μεταβιβάζεται στο κατάντη γενετικό μοντέλο ως αυτόνομα ακουστικά διακριτικά (audio tokens), ενώ

ταυτόχρονα ανακτάται ο αντίστοιχος κβαντισμένος πίνακας (στο διάγραμμα $z_q(x)$ ή $z_q(s)$ βάσει του τύπου) για να τροφοδοτήσει τον αποκωδικοποιητή (Decoder).



Εικόνα 14: Σχήμα περιγραφής του VQ-VAE

Ένα σοβαρό αρχιτεκτονικό πρόβλημα ανακύπτει κατά τη φάση της οπισθοδιάδοσης (backpropagation). Επειδή η συνάρτηση argmin είναι μια διακριτή συνάρτηση βήματος (step function), η μαθηματική παράγωγός της ισούται με μηδέν σε ολόκληρο το πεδίο ορισμού της. Κατά την αναδρομή, οι διαβαθμίσεις του σφάλματος που επιστρέφουν από τον αποκωδικοποιητή προσκρούουν στο επίπεδο κβαντισμού και μηδενίζονται ακαριαία. Η ροή διακόπτεται. Αυτό το φαινόμενο στερεί από το δίκτυο του κωδικοποιητή τη δυνατότητα να ανανεώσει τα βάρη του, καθυλώνοντας την εκπαίδευση. Την άρση αυτού του αδιεξόδου βελτιστοποίησης αναλαμβάνει ο Εκτιμητής Ευθείας Διέλευσης (Straight-Through Estimator - STE), ο οποίος αναπτύχθηκε στο ίδιο αρχιτεκτονικό πλαίσιο από τους [van den Oord et al. \(2017\)](#). Ο STE παρακάμπτει το μη διαφορίσιμο αυτό βήμα κατά το ανάστροφο πέρασμα (backward pass), αντιγράφοντας τις διαβαθμίσεις από την είσοδο του αποκωδικοποιητή z_q απευθείας στον κωδικοποιητή z . Στο επίπεδο της συμβολικής μοντελοποίησης της αρχιτεκτονικής, η ροή αυτή αποδίδεται μαθηματικά ως:

$$z_q \leftarrow z + \text{sg} [z_q - z] \quad (25)$$

Στη μαθηματική αυτή διατύπωση, ο τελεστής $sg[\cdot]$ αντιπροσωπεύει τη διακοπή της διαβάθμισης (stop-gradient). Κατά το εμπρόσθιο πέρασμα (forward pass), ο τελεστής λειτουργεί ως συνάρτηση ταυτότητας, αλλά στο ανάστροφο πέρασμα μπλοκάρει τη ροή των κλίσεων, επιτρέποντας τη σταθερή εκπαίδευση του δικτύου. Η γεωμετρική συμπεριφορά αυτής της παράκαμψης αποτυπώνεται στο δεξί τμήμα της [Εικόνας 14](#). Στον συνεχή χώρο των ενσωματώσεων, η τρέχουσα έξοδος του κωδικοποιητή $z_e(x)$ έλκει την πλησιέστερη διακριτή πρωτότυπη αναπαράσταση e_2 .

Χάρη στον μηχανισμό STE, η κόκκινη διαβάθμιση σφάλματος $\nabla_z L$ παρακάμπτει τη μαθηματική ασυνέχεια του κβαντιστή (όπως φαίνεται από το καμπύλο πορτοκαλί βέλος στο αριστερό διάγραμμα) και εφαρμόζεται απευθείας στην έξοδο του κωδικοποιητή. Η κλίση αυτή «ωθεί» τη θέση της αναπαράστασης στον χώρο, επιτρέποντας την τροποποίηση και τη βελτιστοποίηση της δομής του βιβλίου κωδικών στον επόμενο υπολογιστικό κύκλο (forward pass).

3.2.3 Το Τείχος Κλιμάκωσης του Ρυθμού Μετάδοσης και η Κβαντοποίηση

Υπολειμματικών Διανυσμάτων (RVQ)

Η κλασική κβαντοποίηση διανυσμάτων αποδίδει εξαιρετικά σε ακολουθίες χαμηλής διαστατικότητας, όπως τα διακριτικά ομιλίας (speech tokens), στη βάση του μεθοδολογικού προτύπου των [van den Oord et al. \(2017\)](#). Προσक्रούει, όμως, σε ένα ανυπέρβλητο δομικό τείχος όταν καλείται να κωδικοποιήσει σύνθετα, πολυοργανικά ακουστικά σήματα. Η πιστή καταγραφή των ακουστικών υφών (acoustic textures), των γρήγορων μεταβατικών φαινομένων (transients) και του στερεοφωνικού χώρου απαιτεί υψηλό ρυθμό μετάδοσης δεδομένων (bitrate). Στο απλό VQ, η αύξηση του bitrate προϋποθέτει την επέκταση του μεγέθους του βιβλίου κωδικών K . Όπως τεκμηριώνεται από τον Gray (1984), η χωρητικότητα πληροφορίας του codebook κλιμακώνεται λογαριθμικά, σύμφωνα με τη σχέση $B = \log_2(K)$ bits ανά πλαίσιο. Για την επίτευξη ενός επαγγελματικού bitrate της τάξης των 24 kbps, το βιβλίο κωδικών θα έπρεπε να περιλαμβάνει τρισεκατομμύρια διακριτές καταχωρήσεις. Η απαίτηση αυτή προκαλεί εκθετική συμφόρηση στη μνήμη. Οδηγεί, επίσης,

σε πλήρη κατάρρευση του βιβλίου κωδικών (codebook collapse). Οι [Zeghidour et al. \(2021\)](#) επισημαίνουν ότι σε αυτή την παθολογική κατάσταση το δίκτυο ενεργοποιεί επιλεκτικά ελάχιστα διανύσματα, ενώ οι υπόλοιποι νευρώνες του χώρου αναπαράστασης αδρανούν μόνιμα.

Την κρίση αυτή ξεπερνούν οι σύγχρονες αρχιτεκτονικές με την εφαρμογή της Κβαντοποίησης Υπολειμματικών Διανυσμάτων (Residual Vector Quantization - RVQ), μια προσέγγιση που αναδεικνύεται στο έργο των [Zeghidour et al. \(2021\)](#). Αντί για ένα μοναδικό, γιγαντιαίο βιβλίο κωδικών, η μέθοδος RVQ συνδέει μια αλληλουχία από μικρότερα, ανεξάρτητα βιβλία κωδικών σε μια διαδοχική στοίβα υπολοίπων (residual stack), υιοθετώντας την ιστορική μαθηματική βάση των Juang & Gray (1982). Στα σύγχρονα δίκτυα χρησιμοποιούνται χρησιμοποιούνται 4, 8 ή 16 επίπεδα, με $K=1024$ καταχωρήσεις το καθένα. Στο επίπεδο της αρχιτεκτονικής υλοποίησης, ο μαθηματικός κύκλος της RVQ εκτελείται διαδοχικά και μπορεί να αποδοθεί συμβολικά ως εξής:

1. Ο πρώτος κβαντιστής q_1 επεξεργάζεται το αρχικό συνεχές διάνυσμα z , παράγοντας την πρώτη κβαντισμένη αναπαράσταση:

$$z_q^1 = q_1(z) \quad (26)$$

2. Το σύστημα υπολογίζει το υπολειμματικό διάνυσμα σφάλματος (residual error vector) r_1 , αφαιρώντας την κβαντισμένη τιμή από την αρχική:

$$r_1 = z - z_q^1 \quad (27)$$

3. Αυτό το συνεχές διάνυσμα σφάλματος τροφοδοτείται αμέσως στον κβαντιστή του δεύτερου επιπέδου, αποδίδοντας:

$$z_q^2 = q_2(r_1) \quad (28)$$

Ο συγκεκριμένος βρόχος αφαίρεσης και επανακβαντισμού επαναλαμβάνεται για το σύνολο των N_q σταδίων. Στη τρέχουσα βιβλιογραφία, η μαθηματική γενίκευση για την

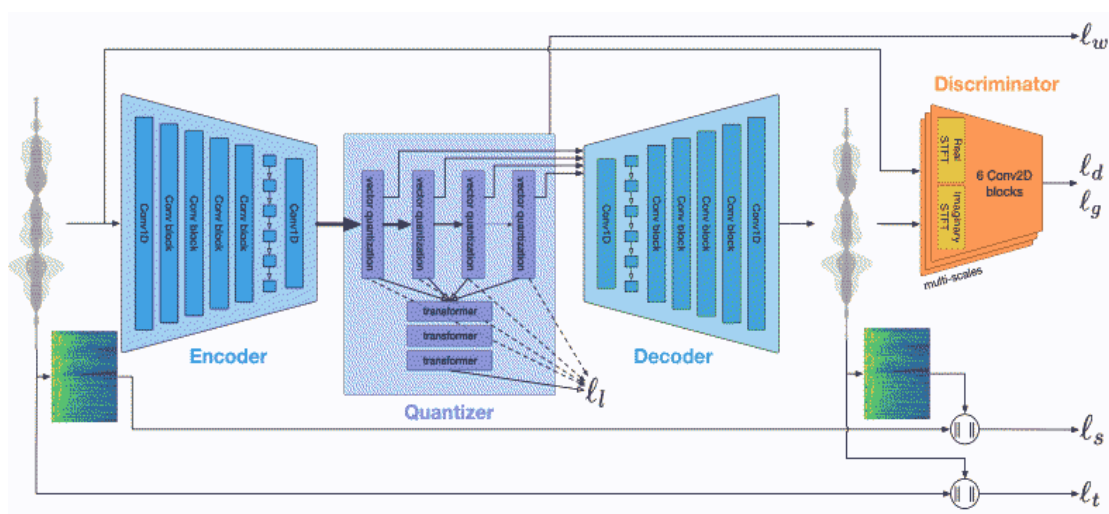
ανακατασκευή του τελικού, πλήρως κβαντισμένου διανύσματος z_q προκύπτει από την απλή γραμμική άθροιση των επιμέρους κβαντισμένων καταχωρήσεων όλων των επιπέδων:

$$z_q = \sum_{j=1}^{N_q} z_q^j \quad (29)$$

Η παραπάνω πολυεπίπεδη δομή κλιμάκωσης αξιοποιείται υποδειγματικά στον προηγμένο νευρωνικό κωδικοποιητή EnCodec (Défossez et al., 2022). Η συνολική υπολογιστική ροή και το σύστημα εκπαίδευσης του μοντέλου οπτικοποιούνται αναλυτικά στην [Εικόνα 15](#). Όπως παρατηρούμε στο διάγραμμα, η αρχιτεκτονική δομείται γύρω από τρία κεντρικά μπλοκ, η σταθερότητα των οποίων διασφαλίζεται από ένα σύνθετο σύνολο συναρτήσεων απώλειας (loss terms):

- **Encoder & Decoder:** Ο κωδικοποιητής μετατρέπει την κυματομορφή Timeline ήχου σε έναν συνεχή λανθάνοντα χώρο μέσω επιπέδων Conv1D, ενώ ο αποκωδικοποιητής την ανακατασκευάζει. Η πιστότητα της ανακατασκευής ελέγχεται απευθείας στο πεδίο του χρόνου μέσω της απώλειας $L_I(l_i)$ και στο πεδίο των συχνοτήτων μέσω της φασματικής απώλειας πολλαπλών κλιμάκων (l_s), η οποία συγκρίνει τα φασματογραφήματα Mel εισόδου και εξόδου.
- **Quantizer Block (RVQ):** Στο κέντρο του διαγράμματος αποτυπώνεται η παράλληλη/διαδοχική διάταξη των κβαντιστών (vector quantization). Κάθε επίπεδο RVQ παράγει διακριτούς δείκτες, οι οποίοι αθροίζονται για να τροφοδοτήσουν τον Decoder. Η διαδικασία αυτή ελέγχεται από την απώλεια δέσμευσης κωδικοποιητή (l_w - commitment loss). Επιπλέον, το μοντέλο εισάγει προαιρετικά ένα μικρό γλωσσικό μοντέλο Transformer (Transformer language model) το οποίο εκτελεί εντροπική κωδικοποίηση (entropy coding) πάνω στους διακριτούς δείκτες, βελτιστοποιώντας την κατανομή τους υπό την επίβλεψη της απώλειας εντροπίας (l_l), γεγονός που μειώνει ακόμη περισσότερο το απαιτούμενο εύρος ζώνης.

- **Multi-Scale STFT Discriminator:** Κατά την εκπαίδευση, το ανακατασκευασμένο σήμα οδηγείται σε έναν Discriminator 6 συνελκτικών επιπέδων (Conv2D), ο οποίος αναλύει το σήμα σε πραγματικό και φανταστικό μέρος μέσω STFT. Το δίκτυο εκπαιδεύεται με αντιπαραθετικές απώλειες για τον δημιουργό (l_g - generator loss) και τον εξομοιωτή (l_d - discriminator loss), εξαλείφοντας τα ακουστικά τεχνουργήματα.



Εικόνα 15 : Αρχιτεκτονική του νευρωνικού κωδικοποιητή ENCODEC

Αυτή η πολυεπίπεδη δομή μεταβάλλει ριζικά την κλιμάκωση της χωρητικότητας, καθώς η πληροφορία αυξάνεται πλέον πολλαπλασιαστικά και όχι εκθετικά. Για παράδειγμα μια στοίβα RVQ 8 επιπέδων με $K=1024$ καταχωρήσεις ανά επίπεδο μπορεί να αναπαραστήσει $1024^8 = 2^{80}$ μοναδικές διαμορφώσεις διανυσμάτων. Για να το πετύχει αυτό, δεσμεύει μόλις $8 \cdot 1024 = 8192$ συνολικές παραμέτρους στη μνήμη, επιτρέποντας στον EnCodec να διατηρεί αναλλοίωτη την ακουστική πιστότητα (acoustic fidelity) της μουσικής, ακόμη και σε συνθήκες ακραίας συμπίεσης.

3.2.4 Πεπερασμένη Βαθμωτή Κβαντοποίηση (FSQ)

Η Κβαντοποίηση Υπολειμματικών Διανυσμάτων (Residual Vector Quantization - RVQ) αυξάνει τη χωρητικότητα πληροφορίας πολλαπλασιαστικά, όπως τεκμηριώνουν οι [Zeghidour et al. \(2021\)](#). Παρά ταύτα, παραμένει στενά συνδεδεμένη με τον μη διαφορίσιμο Εκτιμητή Ευθείας Διέλευσης (Straight-Through Estimator - STE) που πρότειναν οι [van den Oord et al. \(2017\)](#). Επιπλέον, η RVQ απαιτεί επίπονη ρύθμιση υπερπαραμέτρων για την εξισορρόπηση των απωλειών του βιβλίου κωδικών (codebook loss) και της απώλειας δέσμευσης (commitment loss). Προκειμένου να εξαλειφθεί αυτός ο φόρτος βελτιστοποίησης, οι σύγχρονες νευρωνικές ακουστικές αναπαραστάσεις επιστρατεύουν όλο και συχνότερα ένα εναλλακτικό σχήμα διακριτοποίησης: την Πεπερασμένη Βαθμωτή Κβαντοποίηση (Finite Scalar Quantization - FSQ), η οποία αναδείχθηκε από τους [Mentzer et al. \(2023\)](#).

Η FSQ ανατρέπει την κλασική προσέγγιση. Αντί να διοχετεύει τα συνεχή χαρακτηριστικά του κωδικοποιητή σε έναν πίνακα αναζήτησης (lookup table) μέσω κοστοβόρων γεωμετρικών αναζητήσεων πλησιέστερου γείτονα, μετατοπίζει το βάρος στη ντετερμινιστική βαθμωτή στρογγυλοποίηση (scalar rounding) επί ενός σταθερού πλέγματος (grid). Ο κωδικοποιητής προβάλλει αρχικά τη συνεχή αναπαράσταση z_s μέσω ενός γραμμικού επιπέδου σε έναν αυστηρά περιορισμένο, χαμηλής διάστασης χώρο συντεταγμένων R^d , όπου η διάσταση d κυμαίνεται συνήθως μεταξύ 4 και 6. Κάθε διάσταση περιορίζεται μέσω της συνάρτησης υπερβολικής εφαπτομένης $\tanh$ και αντιστοιχίζεται σε ένα διακριτό επίπεδο. Η διαδικασία αυτή της βαθμωτής στρογγυλοποίησης μπορεί να αποδοθεί συμβολικά ως:

$$\hat{z}_i = \text{round} \left(f(z_i) \right) \quad (30)$$

Στη μαθηματική αυτή συνάρτηση, η $f(\cdot)$ κλιμακώνει την οριοθετημένη τιμή εισόδου ώστε να εναρμονιστεί με έναν προκαθορισμένο αριθμό διακριτών βημάτων L_i . Η προκύπτουσα πλειάδα (tuple) των διακριτών δεικτών μετατρέπεται στη συνέχεια απευθείας σε ένα ενιαίο ακέραιο ακουστικό διακριτικό (audio token) μέσω αλλαγής βάσης (base conversion).

Επειδή το πλέγμα κβαντισμού είναι απόλυτα σταθερό και στηρίζεται αποκλειστικά σε ντετερμινιστικές στρογγυλοποιήσεις, η FSQ παρουσιάζει τρία σημαντικά πλεονεκτήματα:

- Δεν απαιτεί εκπαιδευμένες παραμέτρους για βιβλία κωδικών (codebooks), ελαφρύνοντας δραματικά τη μνήμη.
- Εξαλείφει πλήρως τις βοηθητικές απώλειες δέσμευσης (commitment losses) από τη συνολική συνάρτηση απώλειας του δικτύου.
- Παρακάμπτει οριστικά τον κίνδυνο κατάρρευσης του codebook (codebook collapse), καθώς δεν υφίστανται «νεκροί» νευρώνες προς ανανέωση.

Αυτό καθιστά την FSQ μια εξαιρετικά αξιόπιστη εναλλακτική λύση για τις μακροπρόθεσμες αλυσίδες συμπίεσης στα σύγχρονα συστήματα παραγωγής μουσικής από κείμενο, διασφαλίζοντας καθαρή σύγκλιση κατά την εκπαίδευση.

3.2.5 Μετρήσεις Σταθεροποίησης: Κανονικοποίηση Διανύσματος Υπερσφαίρας

Οι τυπικές στοίβες RVQ παρουσιάζουν μια συγκεκριμένη γεωμετρική αδυναμία κατά τη συμπίεση ορχηστρικής μουσικής υψηλής πιστότητας. Οι συνεχείς αναπαραστάσεις που εξέρχονται από τα συνελκτικά επίπεδα κλιμακώνονται συχνά ανομοιόμορφα στις διάφορες συχνότητες. Αυτή η χωρική μετατόπιση (spatial shift) στρεβλώνει την κλασική αναζήτηση πλησιέστερου γείτονα, καθώς η Ευκλείδεια απόσταση τείνει να ευνοεί ένα ελάχιστο υποσύνολο συντεταγμένων του βιβλίου κωδικών, με αποτέλεσμα τη δημιουργία «νεκρών» επιπέδων του βιβλίου κωδικών (dead codebook levels). Σύμφωνα με τον Kumar (2024), την αρχιτεκτονική αυτή ευπάθεια επιλύουν οι σύγχρονοι κωδικοποιητές ήχου υψηλής χωρητικότητας μέσω της Κανονικοποίησης Διανύσματος Υπερσφαίρας (Hypersphere Vector Normalization) όπου όλα τα διανύσματα κλειδώνονται σε μοναδιαίο μήκος με την αναζήτηση να παύει να επηρεάζεται από το μέγεθος των διανυσμάτων και να καθορίζεται αποκλειστικά από τη μεταξύ τους γωνία σε αντίθεση με τον τυπικό RVQ, όπου τα διανύσματα επεκτείνονται ανεξέλεγκτα στον χώρο αναπαράστασης και η ευκλείδεια

απόσταση απομονώνει απομακρυσμένες καταχωρήσεις με τις οποίες τελικά να αδρανούν λόγω έλλειψης ενεργοποίησης.

Πριν από την εισαγωγή των δεδομένων στο μπλοκ ελαχιστοποίησης του πλησιέστερου γείτονα, πραγματοποιείται μια κρίσιμη γεωμετρική προσαρμογή. Στο επίπεδο της αρχιτεκτονικής υλοποίησης, η κανονικοποίηση τόσο του συνεχούς υπολειμματικού διανύσματος του κωδικοποιητή r_{j-1} όσο και όλων των εκπαιδευσιμων καταχωρήσεων του βιβλίου κωδίκων C_j σε μοναδιαία διανύσματα (unit vectors) με βάση την L_2 νόρμα τους αποδίδεται συμβολικά ως:

$$\bar{r}_{j-1} = \frac{r_{j-1}}{\|r_{j-1}\|_2}, \quad \bar{e}_i = \frac{e_i}{\|e_i\|_2} \quad (31)$$

Το βήμα της αναζήτησης για τον κβαντισμό εκτελείται πλέον επάνω σε αυτές τις κανονικοποιημένες τιμές:

$$z_{q,s}^j = \underset{i}{\operatorname{argmin}} \|\bar{r}_{j-1} - \bar{e}_i\|_2^2 \quad (32)$$

Η συγκεκριμένη διαδικασία περιορίζει την αλληλεπίδραση των διανυσμάτων στο ομοιόμορφο κέλυφος μιας πολυδιάστατης υπερσφαίρας. Ως εκ τούτου, η μετρική αναζήτησης του codebook μεταβάλλεται ριζικά: η απόλυτη Ευκλείδεια χωρική απόσταση αντικαθίσταται από τη γωνιακή εγγύτητα συνημίτονου (cosine proximity). Η δομική αυτή βελτίωση εμποδίζει ένα μεμονωμένο, εξαιρετικά ενεργό διάνυσμα του κωδικοποιητή να μονοπωλεί και να υπερφορτώνει τις αναθέσεις του βιβλίου κωδίκων. Το δίκτυο εξαναγκάζεται να αξιοποιεί ομοιόμορφα ολόκληρο τον διαθέσιμο προϋπολογισμό του codebook. Η εκπαίδευση σταθεροποιείται εντυπωσιακά. Παράλληλα, ο κωδικοποιητής αποκτά την ικανότητα να διατηρεί εξαιρετικά λεπτές και απαιτητικές μουσικές λεπτομέρειες, όπως οι σύνθετοι αρμονικοί παλμοί των κυμβάλων (cymbal shimmers) και οι ακαριαίες παροδικές επιθέσεις των κρουστών (transient drum attacks).

3.2.6 Βελτιστοποίηση Βιβλίου Κωδικών

Η καθαρή σύγκλιση των εξόδων του κωδικοποιητή και των στατικών συντεταγμένων του βιβλίου κωδικών (codebook) στον πολυδιάστατο χώρο απαιτεί μια πολυπαραγοντική συνάρτηση κόστους (loss function). Όπως αναλύεται στα θεμελιώδη έργα των [van den Oord et al. \(2017\)](#) και [Zeghidour et al. \(2021\)](#), η συνολική απώλεια βελτιστοποίησης εξισορροπεί τρεις διακριτούς στόχους: την ακρίβεια της ανακατασκευής (reconstruction accuracy), την προσαρμογή των διανυσμάτων του βιβλίου κωδικών και τη σταθεροποίηση του κωδικοποιητή. Η μαθηματική διατύπωση για την ειδική απώλεια του βιβλίου κωδικών (codebook loss) για μια δομή κβαντισμού πολλαπλών σταδίων, τυποποιείται στη σύγχρονη βιβλιογραφία ως:

$$L_{codebook}(\theta_{enc}, C) = \sum_{j=1}^{N_q} \left(\|sg[r_{j-1}] - z_q^j\|_2^2 + \beta \|r_{j-1} - sg[z_q^j]\|_2^2 \right) \quad (33)$$

Όπου:

- $r_0 = z$: Ορίζει την αρχική συνθήκη του υπολοίπου, η οποία ισούται με τη συνεχή λανθάνουσα αναπαράσταση εισόδου.
- β : Αποτελεί την υπερπαραμέτρο δέσμευσης (commitment hyperparameter) που καθορίζει ο χρήστης, με τη βέλτιστη τιμή της να ορίζεται συνήθως στο 0,25.
- sg : Ο τελεστής διακοπής διαβάθμισης (stop-gradient).

Η αρχιτεκτονική αυτή διάταξη επιβάλλει έναν διπλό γεωμετρικό έλεγχο. Το πρώτο σκέλος της εξίσωσης ενημερώνει τα διανύσματα του βιβλίου κωδικών μετατοπίζοντάς τα προς τις εξόδους του κωδικοποιητή. Αντίθετα, οι δημιουργοί επισημαίνουν ότι το δεύτερο σκέλος — το οποίο συνιστά την απώλεια δέσμευσης (commitment loss)— εξαναγκάζει τις εξόδους του κωδικοποιητή να «δεσμευτούν» στις επιλεγμένες συντεταγμένες του βιβλίου κωδικών, εμποδίζοντάς τις να διακυμαίνονται ανεξέλεγκτα κατά την εκπαιδευτική διαδικασία.

Μετά από μια αρχική φάση εκπαίδευσης, το σύστημα παγώνει τα βιβλία κωδικών. Παράλληλα, εισάγει δίκτυα απώλειας βαθιάς διακριτικής ικανότητας (discriminator loss). Οι κωδικοποιητές προσφέρουν πλέον έναν εξαιρετικά στιβαρό, δομημένο χώρο διακριτών συμβόλων (discrete token space). Όπως καταδεικνύεται στο μεθοδολογικό πλαίσιο των [Copet et al. \(2023\)](#), η προσέγγιση αυτή εισάγει μια ριζική αλλαγή παραδείγματος κατά την παραγωγική διαδικασία. Όταν το βαθύτερο μοντέλο μετατροπής κειμένου σε ήχο τίθεται σε λειτουργία, αντιμετωπίζει αυτά τα επίπεδα RVQ ως έναν στόχο πολλαπλών επιπέδων (multi-codebook target). Το μοντέλο απαλλάσσεται από την ανάγκη να προβλέψει μια ενιαία, χαοτική χρονική ακολουθία συνεχών αριθμών. Αντίθετα, εκπαιδεύεται στην παραγωγή ενός παράλληλου πλέγματος συγχρονισμένων διακριτών δεικτών. Αυτή η δομική οργάνωση επιτρέπει τη σταθερή μετάφραση των αφηρημένων γλωσσικών προτροπών σε λεπτομερείς, πολυεπίπεδες ακουστικές και μουσικές εννοχρηστώσεις, διασφαλίζοντας υψηλή πιστότητα κατά τη σύνθεση.

3.3 Διατροπική Δομική Χαρτογράφηση (Transformer/Attention Core)

Μόλις η προτροπή κειμένου μετατραπεί σε συνεχή σημασιολογικά διανύσματα (Ενότητα 3.1) και ο ηχητικός χώρος-στόχος συμπιέζεται σε διαχειρίσιμες λανθάνουσες μεταβλητές ή διακριτούς δείκτες βιβλίου κωδικών (Ενότητα 3.2), το σύστημα πρέπει να εκτελέσει την κρίσιμη εργασία της διατροπικής δομικής χαρτογράφησης. Αυτή η φάση χειρίζεται ο πυρήνας του δικτύου, ο οποίος χρησιμοποιεί εξειδικευμένους μηχανισμούς προσοχής και προσαρμοστικές τοπολογίες κανονικοποίησης για να υπολογίσει τις ακριβείς αριθμητικές εξαρτήσεις μεταξύ των διακριτικών γλωσσικής εξαρτημένης επεξεργασίας και των διακριτικών ακουστικής εξόδου.

3.3.1 Μηχανική Συνθήκης Διασταυρούμενης Προσοχής

Η βασική αρχιτεκτονική για την άμεση αντιστοίχιση μιας πληροφορίας συνθήκης (κείμενο) σε μια ακολουθία-στόχο (ήχος) είναι ο μηχανισμός της διασταυρούμενης προσοχής (cross-attention), ο οποίος εισήχθη στο έργο του [Vaswani \(2017\)](#). Σε αντίθεση με την τυπική αυτοπροσοχή (self-attention), όπου τα ερωτήματα (queries), τα κλειδιά (keys) και οι τιμές (values) προέρχονται από την ίδια ακριβώς ακολουθία εισόδου, η διασταυρούμενη προσοχή λειτουργεί ως μια πληροφοριακή γέφυρα ανάμεσα σε δύο εντελώς διαφορετικούς διανυσματικούς χώρους.

Όπως φαίνεται και στο συνολικό αρχιτεκτονικό διάγραμμα στην [Εικόνα 7](#), ο μηχανισμός αυτός αποτελεί τον συνδετικό κρίκο μεταξύ των δύο βασικών μπλοκ του Transformer:

- **Η Πλευρά της Συνθήκης (Encoder - Αριστερά):** Αντιπροσωπεύει τον κωδικοποιητή κειμένου. Επεξεργάζεται την κειμενική προτροπή και εξάγει τις τελικές σημασιολογικές αναπαραστάσεις. Από την κορυφή αυτού του μπλοκ, η πληροφορία διακλαδίζεται και στέλνει δύο βέλη προς το δεξί τμήμα, μεταφέροντας τα Κλειδιά (K) και τις Τιμές (V).
- **Η Πλευρά της Παραγωγής (Decoder - Δεξιά):** Αντιπροσωπεύει τη γενετική μηχανή του ήχου. Στο μέσο κάθε επιπέδου του αποκωδικοποιητή, αναπτύσσεται το επίπεδο Multi-Head Attention της Διασταυρούμενης Προσοχής. Το επίπεδο αυτό δέχεται τα Ερωτήματα (Q) από τις χαμηλότερες ακουστικές στρώσεις του Decoder και ταυτόχρονα υποδέχεται τα βέλη K και V από τον Encoder, επιτυγχάνοντας τον διατροπικό συγκερασμό.

Έστω $H_{text} \in \mathbb{R}^{L \times D_{text}}$ ο συνεχής πίνακας χαρακτηριστικών που εξάγεται από την προτροπή κειμένου μέσω του κωδικοποιητή front-end, όπου η μεταβλητή L ορίζει το μήκος της ακολουθίας κειμένου. Αντίστοιχα, έστω $H_{audio} \in \mathbb{R}^{S \times D_{audio}}$ η τρέχουσα κατάσταση της ηχητικής ακολουθίας εντός του γενετικού δικτύου, όπου η μεταβλητή S αντιπροσωπεύει το συμπιεσμένο χρονικό μήκος πλαισίου (frame length) του ακουστικού σήματος. Το επίπεδο

διασταυρούμενης προσοχής προβάλλει αυτούς τους ξεχωριστούς πίνακες σε διακριτούς λειτουργικούς χώρους:

$$\mathbf{Q} = \mathbf{H}_{audio} \mathbf{W}_Q \quad (\text{Ερωτήματα που προέρχονται από τον ήχο}) \quad (34)$$

$$\mathbf{K} = \mathbf{H}_{text} \mathbf{W}_K \quad (\text{Κλειδιά που προέρχονται από το κείμενο}) \quad (35)$$

$$\mathbf{V} = \mathbf{H}_{text} \mathbf{W}_V \quad (\text{Τιμές που προέρχονται από το κείμενο}) \quad (36)$$

όπου

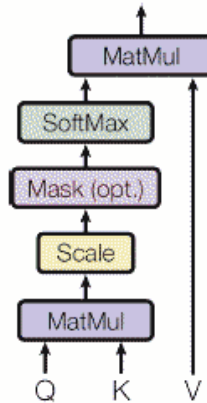
$$\mathbf{W}_Q \in \mathbb{R}^{D_{audio} \times d_k}, \quad \mathbf{W}_K \in \mathbb{R}^{D_{text} \times d_k}, \quad \text{και} \quad \mathbf{W}_V \in \mathbb{R}^{D_{text} \times d_v} \quad (37)$$

είναι εκπαιδεύσιμοι πίνακες γραμμικών βαρών. Στη συνέχεια, ο υπολογισμός της κλιμακωτής προσοχής εσωτερικού γινομένου εκτελείται επάνω σε αυτούς τους διατροπικούς τανυστές (tensors) (Εικόνα 16):

$$\mathbf{A}_{cross} = \text{softmax} \left(\frac{\mathbf{Q} \mathbf{K}^T}{\sqrt{d_k}} \right) \in \mathbb{R}^{S \times L} \quad (38)$$

$$\mathbf{H}_{context} = \mathbf{A}_{cross} \mathbf{V} \in \mathbb{R}^{S \times d_v} \quad (39)$$

Ο πίνακας βαρών διασταυρούμενης προσοχής $\mathbf{A}_{cross}$ αντιστοιχίζει κάθε χρονικό πλαίσιο του ηχητικού σήματος (S) έναντι κάθε σημασιολογικού διακριτικού (subword token) της κειμενικής προτροπής (L).



Εικόνα 16: Scaled Dot-Product Attention

Με βάση την [Εικόνα 16](#), η ροή εκτελείται από κάτω προς τα πάνω:

1. **MatMul (Πρώτο επίπεδο):** Εκτελείται ο πολλαπλασιασμός του πίνακα ερωτημάτων του ήχου (Q) με τον αναστραμμένο πίνακα κλειδιών του κειμένου (K^T), παράγοντας τις ακατέργαστες βαθμολογίες συγγένειας.
2. **Scale:** Οι βαθμολογίες διαιρούνται με τον παράγοντα κλιμάκωσης $\sqrt{d_k}$ για τη σταθεροποίηση των κλίσεων.
3. **Mask (opt.):** Στο πλαίσιο της διασταυρούμενης προσοχής, το βήμα αυτό παρακάμπτεται (Bypassed), καθώς ολόκληρο το κειμενικό περιβάλλον είναι εκ των προτέρων γνωστό και διαθέσιμο στο μοντέλο.
4. **SoftMax & MatMul (Τελικό επίπεδο):** Η συνάρτηση SoftMax μετατρέπει τις κλιμακωμένες τιμές σε μια κανονικοποιημένη κατανομή πιθανοτήτων (πίνακας A_{cross}), η οποία πολλαπλασιάζεται με τον πίνακα τιμών του κειμένου (V) για την εξαγωγή της τελικής αναπαράστασης περιβάλλοντος ($H_{context}$).

Ο πίνακας βαρών διασταυρούμενης προσοχής A_{cross} που παράγεται αντιστοιχίζει κάθε χρονικό πλαίσιο του ηχητικού σήματος (S) έναντι κάθε σημασιολογικού διακριτικού (subword token) της κειμενικής προτροπής (L). Η λειτουργία αυτή γίνεται καλύτερα κατανοητή μέσα από ένα συγκεκριμένο παράδειγμα:

- Εάν η προτροπή κειμένου περιέχει το διακριτικό «παραμόρφωση(εννοώντας του ήχου)», τα γραμμικά βάρη προβολής W_K μεταφράζουν αυτό το σημασιολογικό στοιχείο σε ένα ξεχωριστό, χαρακτηριστικό προφίλ κλειδιού.
- Καθώς η γενετική μηχανή κατασκευάζει ή αφαιρεί επαναληπτικά τον θόρυβο (denoising) από ένα ηχητικό πλαίσιο που παρουσιάζει κορεσμό αρμονικών υψηλής συχνότητας, το αντίστοιχο προφίλ ερωτήματός του (Q) θα καταγράψει μια ιδιαίτερα υψηλή βαθμολογία ομοιότητας (dot-product score) με το συγκεκριμένο κλειδί του κειμένου.

Στη συνέχεια, το επίπεδο της συνάρτησης softmax κανονικοποιεί αυτές τις βαθμολογίες συγγένειας σε σαφείς πιθανότητες. Όπως επισημαίνεται στα έργα των [Liu et al. \(2023\)](#) και [Copet et al. \(2023\)](#), αυτός ο μαθηματικός περιορισμός εξαναγκάζει το δίκτυο να δρομολογήσει τις πληροφορίες πλαισίου από τον πίνακα τιμών του κειμένου (V) απευθείας στην ενεργή διαδρομή παραγωγής του ήχου. Με τον τρόπο αυτό, το κείμενο καθοδηγεί και διαμορφώνει την ακουστική υφή της σύνθεσης με ακρίβεια χρονικού πλαισίου.

3.3.2 Διατροφική Χαρτογράφηση σε Μοντέλα Λανθάνουσας Διάχυσης (Latent Diffusion)

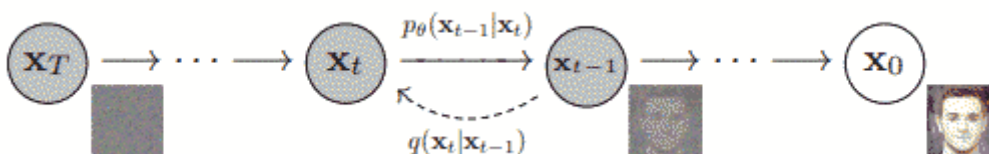
Στα δίκτυα διάχυσης κειμένου σε ήχο (text-to-audio diffusion networks), με χαρακτηριστικό παράδειγμα το μοντέλο AudioLDM που αναπτύχθηκε από τους [Liu et al. \(2023\)](#), το στρώμα της διατροφικής δομικής χαρτογράφησης ενσωματώνεται απευθείας μέσα στα κρυφά μπλοκ (hidden blocks) μιας χωροχρονικής νευρωνικής ραχοκοκαλιάς (backbone). Η αρχιτεκτονική αυτή βάση υλοποιείται συνήθως μέσω ενός δικτύου U-Net, στη βάση του δομικού προτύπου των [Ronneberger et al. \(2015\)](#), ή εναλλακτικά μέσω ενός Μετασχηματιστή Διάχυσης (Diffusion Transformer - DiT), σύμφωνα με τη μεθοδολογία των [Peebles & Xie \(2023\)](#).

Όπως εισηγήθηκαν οι [Ho et al. \(2020\)](#), η εισαγωγή της κειμενικής συνθήκης στη γενετική τροχιά της διάχυσης ακολουθεί μια αυστηρά καθορισμένη στοχαστική πορεία, η οποία οπτικοποιείται στην [Εικόνα 17](#). Παρά το γεγονός ότι το αρχικό θεωρητικό πλαίσιο των [Ho et al. \(2020\)](#) αναπτύχθηκε στον τομέα της υπολογιστικής όρασης (όπως υποδηλώνει η σχηματική αναπαράσταση της ανθρώπινης μορφής στο διάγραμμα), η μαθηματική και γεωμετρική συμπεριφορά του μοντέλου παραμένει πανομοιότυπη κατά τη διαχείριση ηχητικών δεδομένων. Στο πλαίσιο αυτό, οι μεταβλητές x_0 , x_b , x_T αντιπροσωπεύουν τις καταστάσεις του λανθάνοντος φασματογραφήματος ήχου κατά μήκος του άξονα του χρόνου διάχυσης (diffusion time steps):

- **Η Φάση της Αλλοίωσης (Forward Process- $q(x_t | x_{t-1})$):** Αντιπροσωπεύει τη διακεκομμένη τροχιά από αριστερά προς τα δεξιά. Κατά την εκπαίδευση, η καθαρή ακουστική αναπαράσταση (x_0) αλλοιώνεται προοδευτικά και ελεγχόμενα μέσω της

διαδοχικής προσθήκης λευκού θορύβου Gauss. Στο τελικό χρονικό βήμα T , το σήμα έχει μετατραπεί πλήρως σε έναν χαοτικό, μη δομημένο θόρυβο ($x_T \sim \mathcal{N}(0, I)$).

- **Η Φάση της Παραγωγής (Reverse Process – $p_\theta(x_{t-1} | x_t)$):** Αντιπροσωπεύει τη συνεχή τροχιά από δεξιά προς τα αριστερά. Κατά τη δειγματοληψία (sampling), το μοντέλο ξεκινά από τον καθαρό θόρυβο x_T και εκπαιδεύεται να εκτελεί βήμα-βήμα την αφαίρεση του θορύβου (denoising), προσεγγίζοντας σταδιακά την καθαρή δομή της ακουστικής εξόδου x_T .
- **Η Φάση της Εξάρτησης (Conditioning):** Κατά την αντίστροφη αυτή πορεία, ο θορυβώδης τανυστής ήχου x_t διοχετεύεται στον γενετικό πυρήνα (U-Net ή DiT), συνοδευόμενος παράλληλα από τις αναπαραστάσεις της εξωτερικής συνθήκης. Στην τρέχουσα βιβλιογραφία, η συμβολική μοντελοποίηση αυτής της εισόδου τυποποιείται είτε ως ένα καθολικό διάνυμα επεξεργασίας κειμένου z_{text} (μοντέλα CLAP) είτε ως μια ακολουθία πλούσιων κειμενικών ενσωματώσεων H_{text} (μοντέλα T5), οι οποίες εξάγονται από τους προεκπαιδευμένους γλωσσικούς κωδικοποιητές.



Εικόνα 17: Φάσεις Αποθορυβοποίησης

Μέσα σε κάθε συνελικτικό (convolutional) ή προσεκτικό (attentional) μπλοκ του μηχανισμού αποθορυβοποίησης, τα επίπεδα διασταυρούμενης προσοχής (cross-attention layers) αναλαμβάνουν την επεξεργασία των γλωσσικών χαρακτηριστικών με βάση τη μαθηματική τυποποίηση του [Vaswani \(2017\)](#). Στο πλαίσιο αυτό, τα διακριτικά του κειμένου λειτουργούν ως ένα ενεργό φίλτρο συντεταγμένων (coordinate filter). Καθοδηγούν το δίκτυο τροποποιώντας δυναμικά τα επίπεδα πρόβλεψης του θορύβου που πρέπει να αφαιρεθεί σε κάθε χρονικό βήμα της αντίστροφης διαδικασίας. Αυτός ο εσωτερικός γεωμετρικός έλεγχος

επιτρέπει στο μοντέλο να αντιστοιχίζει τις αφηρημένες γλωσσικές εισόδους απευθείας στην εντοπισμένη δημιουργία σύνθετων και δομημένων σχημάτων επάνω στο λανθάνον φασματογράφημα (latent spectrogram). Οι χωρικές αυτές αναπαραστάσεις, κατά την έξοδο τους από τον γενετικό πυρήνα, αποκωδικοποιούνται στη συνέχεια από το backend του σήματος, μετατρέπόμενες σε υψηλής πιστότητας ακουστικές υφές και κυματομορφές.

3.3.3 Διαμορφώσεις Καθολικής Προετοιμασίας: Προσαρμοστική Κανονικοποίηση Επιπέδου (adaLN-Modulation)

Ενώ τα επίπεδα διασταυρούμενης προσοχής (cross-attention layers) που βασίζονται σε διακριτικά στοιχεία υπερέχουν στην εκτέλεση εντοπισμένης και χρονικά ευθυγραμμισμένης αντιστοίχισης χαρακτηριστικών, παρουσιάζουν ένα σημαντικό υπολογιστικό εμπόδιο όταν καλούνται να επιβάλουν ολιστικούς περιορισμούς μακρο-επιπέδου. Τέτοιοι περιορισμοί περιλαμβάνουν τις παραμέτρους του παγκόσμιου ακουστικού περιβάλλοντος, το στυλ πιστότητας της εγγραφής ή το γενικό μουσικό είδος (macro-genre). Για την ενσωμάτωση αυτής της καθολικής προετοιμασίας (global conditioning) χωρίς την υπολογιστική επιβάρυνση που συνεπάγεται η προσοχή πάνω σε πυκνές ακολουθίες, οι σύγχρονες ακουστικές ραχοκοκαλιές τύπου Μετασχηματιστή Διάχυσης (Diffusion Transformer - DiT) επιστρατεύουν όλο και συχνότερα το πλαίσιο της Προσαρμοστικής Κανονικοποίησης Επιπέδου (Adaptive Layer Normalization - adaLN-Modulation), όπως αυτό εισηγήθηκαν οι [Peebles & Xie \(2023\)](#).

Αντί να προβάλλονται οι κειμενικές ενσωματώσεις αποκλειστικά στους χώρους των κλειδιών (keys) και των τιμών (values), ένα καθολικά συγκεντρωμένο διάνυσμα προετοιμασίας κειμένου z_{text} διοχετεύεται μέσω ενός εξειδικευμένου πολυστρωματικού perceptron (Multi-Layer Perceptron - MLP).

Το βοηθητικό δίκτυο αυτό αναλαμβάνει να προβλέψει ένα σύνολο παραμέτρων κλιμάκωσης (scaling) γ και μετατόπισης (shift) β για κάθε διάσταση, παράλληλα με κατάλληλους παράγοντες πύλης (gating factors) α . Στο επίπεδο της αρχιτεκτονικής υλοποίησης, η μαθηματική γενίκευση της δυναμικής αυτής διαμόρφωσης στο εσωτερικό ενός μπλοκ

Transformer (Transformer block) τυποποιείται στη σύγχρονη βιβλιογραφία μέσω του ακόλουθου συστήματος εξισώσεων:

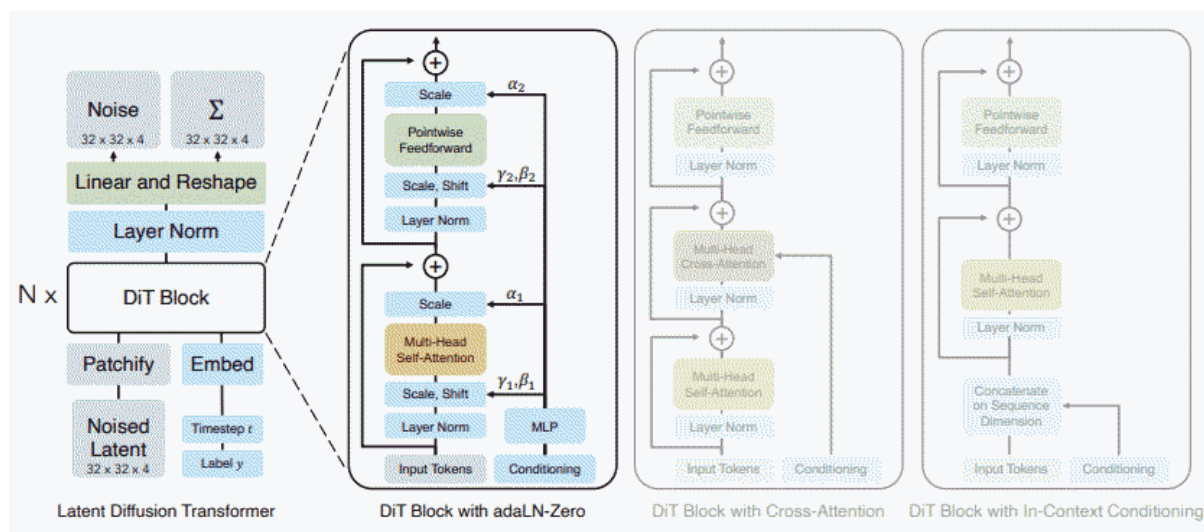
$$adaLN(\mathbf{h}_{audio}, \mathbf{z}_{text}) = \gamma_l(\mathbf{z}_{text}) \odot LayerNorm(\mathbf{h}_{audio}) + \beta_l(\mathbf{z}_{text}) \quad (40)$$

$$\mathbf{h}_{block} = \mathbf{h}_{audio} + \alpha(\mathbf{z}_{text}) \cdot MultiHeadAttention(adaLN(\mathbf{h}_{audio}, \mathbf{z}_{text})) \quad (41)$$

Στη μαθηματική αυτή διατύπωση, ο τελεστής $\odot$ δηλώνει το γινόμενο κατά στοιχεία (element-wise product) και η μεταβλητή $\mathbf{h}_{audio}$ αντιπροσωπεύει την τρέχουσα λανθάνουσα κατάσταση του ήχου. Η υπολογιστική ροή αυτής της διαδικασίας απεικονίζεται λεπτομερώς στο αρχιτεκτονικό διάγραμμα στην [Εικόνα 18](#). Όπως παρατηρούμε η δομή αναλύεται από αριστερά προς τα δεξιά, ξεδιπλώνοντας τη λειτουργία του μοντέλου:

- **Ο Καθολικός Αγωγός (Αριστερό Τμήμα):** Ο θορυβώδης λανθάνων τανυστής (Noised Latent) τεμαχίζεται σε διακριτά χωροχρονικά αποκόμματα (Patchify), ενσωματώνεται γραμμικά (Embed) και τροφοδοτείται σε μια στοίβα από N διαδοχικά μπλοκ DiT, καθοδηγούμενος από το χρονικό βήμα διάχυσης (Timestep t) και την κειμενική ετικέτα (Label y).
- **Το Μπλοκ DiT με adaLN-Zero (Κεντρικό Τμήμα):** Αποτελεί την οπτικοποίηση των παραπάνω εξισώσεων. Η κειμενική συνθήκη (Conditioning) περνά από το επίπεδο MLP και παράγει τις παραμέτρους διαμόρφωσης. Στο διάγραμμα ο μηχανισμός αναφέρεται συγκεκριμένα ως adaLN-Zero, καθώς οι παράγοντες πύλης (α_l, α_2) αρχικοποιούνται αυστηρά στο μηδέν (zero-initialized). Η στρατηγική αυτή είναι κρίσιμη: απομονώνει αρχικά τα υπολειμματικά μπλοκ κατά την έναρξη της εκπαίδευσης, λειτουργώντας ως συνάρτηση ταυτότητας, γεγονός που σταθεροποιεί δραστικά τη ροή των κλίσεων.
- **Εναλλακτικές Διαμορφώσεις (Δεξί Τμήμα):** Για λόγους πληρότητας, οι δημιουργοί παραθέτουν τις παραδοσιακές μεθόδους εισαγωγής συνθήκης, όπως η Διασταυρούμενη Προσοχή (DiT Block with Cross-Attention) και η απευθείας

παράθεση στον άξονα των διακριτικών (In-Context Conditioning). Τα πειραματικά δεδομένα των [Peebles & Xie \(2023\)](#) κατέδειξαν ότι η adaLN-Zero υπερέχει καθαρά, προσφέροντας τη χαμηλότερη τιμή Fréchet απόσταση και βέλτιστη υπολογιστική απόδοση.



Εικόνα 18: Η αρχιτεκτονική Diffusion Transformer (DiT)

Μέσω αυτού του βρόχου διαμόρφωσης (modulation loop), το καθολικό σημασιολογικό σχέδιο παρεμβαίνει άμεσα και μεταβάλλει τις βασικές στατιστικές κατανομές χαρακτηριστικών των ακουστικών επιπέδων. Αυτή η προσέγγιση αναγκάζει ολόκληρο τον χωροχρονικό καμβά να συμμορφωθεί ομοιόμορφα με το επιλεγμένο ακουστικό στυλ κατά μήκος όλων των χρονικών οριζόντων (για παράδειγμα, διατηρώντας ένα σταθερό ύφος «lo-fi βινυλίου της δεκαετίας του 1930»). Η διαδικασία αυτή επιτυγχάνεται άμεσα, παρακάμπτοντας πλήρως τις δαπανηρές υπολογιστικά εξαρτήσεις της αναζήτησης διακριτικών ανά χρονικό βήμα, εξασφαλίζοντας έτσι βελτιστοποιημένη υπολογιστική απόδοση κατά τη φάση της παραγωγής.

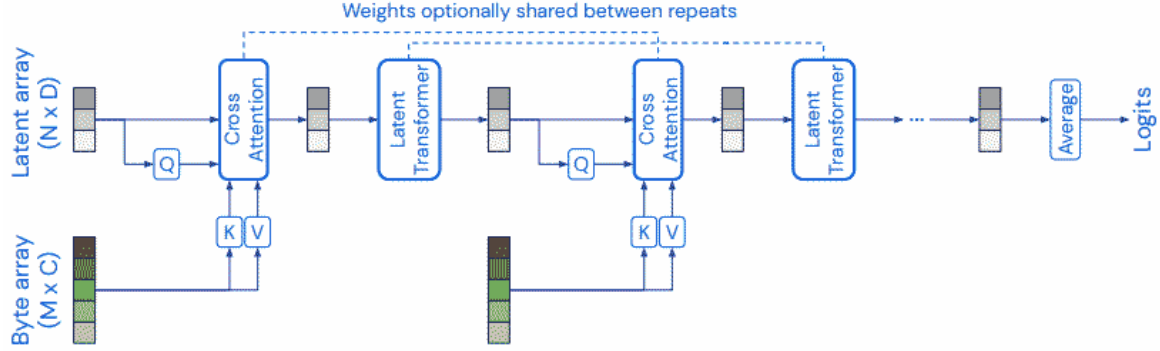
3.3.4 Τυποποίηση Μήκους Ακολουθίας: Αναδειγματοληψία Αντιλήπτη (Perceiver Resampling)

Όπως αναλύθηκε στη διαδικασία γλωσσικής επεξεργασίας (Ενότητα 3.3), η χρήση προηγμένων μοντέλων βαθιάς γλώσσας ή οι τεχνικές επαύξησης λεζαντών μέσω LLMs έχουν ως αποτέλεσμα την παραγωγή κειμενικών ακολουθιών με εξαιρετικά μεταβλητό μήκος L . Οι λεπτομερείς και μακροσκελείς προτροπές χρήστη (prompts) συχνά μεταφράζονται σε εκατοντάδες διακριτικά υπολέξεων (subword tokens). Η απευθείας διοχέτευση αυτών των μη περιορισμένων μητρών κειμένου στα διαδοχικά μπλοκ διασταυρούμενης προσοχής (cross-attention) της ακουστικής μηχανής προκαλεί έντονες διακυμάνσεις στις υπολογιστικές απαιτήσεις. Το γεγονός αυτό διαταράσσει και υπονομεύει τη σταθερή βελτιστοποίηση της μνήμης των τανυστών (tensors).

Προκειμένου να επιβληθεί ένας αυστηρός, ντετερμινιστικός υπολογιστικός προϋπολογισμός, τα σύγχρονα συστήματα μετατροπής κειμένου σε ήχο και μουσική εισάγουν μια ενότητα Αναδειγματοληψίας Αντιλήπτη (Perceiver Resampler). Όπως αναλύεται στο θεμελιώδες έργο του [Jaegle \(2021\)](#), η ενότητα αυτή λειτουργεί ως μια στιβαρή δομική γέφυρα ανάμεσα στις δύο διαφορετικές τροπικότητες, η βασική υπολογιστική φιλοσοφία της οποίας αποτυπώνεται στην [Εικόνα 19](#), όπου στην αρχιτεκτονική Perceiver η ροή οργανώνεται μέσω δύο παράλληλων πινάκων:

- **Ο Πίνακας Εισόδου (Byte Array - $M \times C$):** Αντιπροσωπεύει τον μεταβλητό πίνακα κειμένου του front-end (αντιστοιχεί στη διάσταση L του κειμένου μας). Αυτός ο πίνακας τροφοδοτείται αποκλειστικά στις εισόδους των Κλειδιών (K) και των Τιμών (V) των επιπέδων διασταυρούμενης προσοχής.
- **Ο Λανθάνων Πίνακας (Latent Array - $N \times D$):** Αντιπροσωπεύει τη σταθερή τράπεζα των μαθημένων ερωτημάτων (αντιστοιχεί στη διάσταση $M=64$ του κειμένου μας). Ο πίνακας αυτός εισάγει τα Ερωτήματα (Q) στο πρώτο επίπεδο Cross Attention.
- **Η Επαναληπτική Επεξεργασία (Iterative Processing):** Όπως φαίνεται στη ροή της εικόνας, η έξοδος της διασταυρούμενης προσοχής, η οποία έχει πλέον «κλειδωμένες» σταθερές διαστάσεις $N \times D$, περνά μέσα από μια βαθιά στοίβα υπολογιστικών μπλοκ

αυτοπροσοχής (**Latent Transformer**). Αυτός ο εσωτερικός βρόχος επιτρέπει στα συμπυκνωμένα διακριτικά να αλληλοεπιδράσουν έντονα μεταξύ τους, εκτελώντας μια βαθιά σημασιολογική απόσταξη (semantic distillation) του περιβάλλοντος κειμένου.



Εικόνα 19: Αρχιτεκτονική ροή του Perceiver με τη χρήση επαναληπτικής προσοχής

Ο Αναδειγματολήπτης Αντιλήπτη αρχικοποιεί μια στατική, έντονα περιορισμένη τράπεζα εκπαιδευσιμων λανθανόντων διανυσμάτων ερωτήματος (latent query vectors) $Q_{latent} \in \mathbb{R}^{M \times D}$. Στη δομή αυτή, η μεταβλητή M αποτελεί μια σταθερή, μικρή τιμή (συνήθως $M = 64$ tokens), η οποία είναι πλήρως ανεξάρτητη από το αρχικό μήκος του κειμένου L . Αυτά τα μαθημένα ερωτήματα εκτελούν μια τοπική διεργασία διασταυρούμενης προσοχής πάνω στον πίνακα κειμένου υψηλής διάστασης H_{text} , ο οποίος με τη σειρά του αναλαμβάνει τον ρόλο των Κλειδιών (K) και των Τιμών (V). Στο επίπεδο της αρχιτεκτονικής υλοποίησης, η μαθηματική γενίκευση αυτής της διαδικασίας κλιμακωτής προσοχής εναρμονίζεται με το κλασικό πρότυπο του [Vaswani \(2017\)](#) και τυποποιείται στη σύγχρονη βιβλιογραφία ως:

$$H_{standard} = \text{softmax} \left(\frac{Q_{latent} (H_{text} W_K)^T}{\sqrt{d_k}} \right) (H_{text} W_V) \quad (42)$$

Η προκύπτουσα έξοδος $H_{standard} \in \mathbb{R}^{M \times D}$ λειτουργεί ως μια εξαιρετικά συμπυκνωμένη, τυποποιημένη σημασιολογική απόσταξη (semantic distillation) της αρχικής προτροπής.

Εξαναγκάζοντας όλους τους μεταβλητούς γλωσσικούς περιγραφείς να διέλθουν μέσα από αυτό το στενό ακουστικό σημείο συμφόρησης σταθερού μεγέθους, αρχιτεκτονικές αιχμής όπως το MusicLM των [Agostinelli et al. \(2023\)](#) απομονώνουν με επιτυχία τα θεμελιώδη σημασιολογικά χαρακτηριστικά. Παράλληλα, η συγκεκριμένη προσέγγιση εγγυάται ότι κάθε επόμενο επίπεδο στον κύριο γενετικό πυρήνα εκτελείται υπό απόλυτα ομοιόμορφους, προ-εκχωρημένους περιορισμούς μνήμης VRAM, εξαλείφοντας τον κίνδυνο σφαλμάτων υπερχείλισης (Out-of-Memory).

3.3.5 Διατροπική Χαρτογράφηση σε Αυτοπαλίνδρομα Μοντέλα (Διαπλοκή Ακολουθιών)

Σε αντίθεση με τα μοντέλα λανθάνουσας διάχυσης (latent diffusion models), τα αυτοπαλίνδρομα πλαίσια (autoregressive frameworks), όπως το MusicGen των [Copet et al. \(2023\)](#), αντιμετωπίζουν τη χαρτογράφηση του κειμένου σε μουσική ως ένα ενιαίο πρόβλημα συμπλήρωσης ακολουθίας (sequence completion problem). Καθώς η Κβαντοποίηση Υπολειμματικών Διανυσμάτων (Residual Vector Quantization - RVQ) παράγει ένα παράλληλο πλέγμα πολλαπλών διακριτών δεικτών ανά ηχητικό πλαίσιο, το μοντέλο καλείται να επιλύσει μια σημαντική αρχιτεκτονική πρόκληση: να αντιστοιχίσει μια μεμονωμένη, γραμμική ακολουθία κειμένου σε πολλαπλές, παράλληλες ροές ακουστικών διακριτικών (multi-stream audio tokens), μια διαδικασία που αναλύεται διεξοδικά στα έργα των [Zeghidour et al. \(2021\)](#) και [Défossez et al. \(2022\)](#).

Προκειμένου να επιτευχθεί αυτή η διατροπική σύγκλιση χωρίς να προκληθεί δομική κατάρρευση ή ασυγχρονισμός, τα μοντέλα αυτά εφαρμόζουν μια στρατηγική ενοποιημένης ισοπέδωσης και διαπλοκής. Τα διακριτά αναγνωριστικά (IDs) του κειμένου τοποθετούνται στην αρχή της υπολογιστικής αλυσίδας, λειτουργώντας ως ένα πρόθεμα υπό όρους (conditional prefix). Στη συνέχεια, οι παράλληλες ροές των ακουστικών δεικτών από τα διάφορα επίπεδα του βιβλίου κωδικών RVQ ισοπεδώνονται (flattened) ή διαπλέκονται (interleaved) κατά μήκος της χρονικής διάστασης, ακολουθώντας εξειδικευμένα μοτίβα μετατόπισης ή καθυστέρησης (codebook delay patterns).

Όταν αυτή η ενιαία, συνδυασμένη ροή δεδομένων διοχετεύεται στα τυπικά μπλοκ αυτοπροσοχής πολλαπλών κεφαλών (Multi-Head Self-Attention layers), στη βάση του κλασικού προτύπου των [Vaswani et al. \(2017\)](#), η αρχιτεκτονική δεν απαιτεί ξεχωριστό μηχανισμό διασταυρούμενης προσοχής (cross-attention). Αντίθετα, ο πίνακας αυτοπροσοχής αναλαμβάνει εξολοκλήρου να υπολογίσει τις στατιστικές και αριθμητικές εξαρτήσεις κατά μήκος ολόκληρης της ενοποιημένης ακολουθίας. Επειδή το γλωσσικό πρόθεμα (text prefix) προηγείται χρονολογικά και υπολογιστικά, η αιτιώδης μάσκα κάλυψης (causal mask) επιτρέπει σε κάθε επόμενο ακουστικό διακριτικό —ανεξάρτητα από το επίπεδο RVQ στο οποίο ανήκει— να «κοιτάξει» αποκλειστικά προς τα πίσω στον χρόνο. Με αυτόν τον τρόπο, το δίκτυο υπολογίζει την άμεση μαθηματική και σημασιολογική του εξάρτηση από τις αρχικές γλωσσικές οδηγίες του χρήστη. Αυτή η ενοποιημένη στρατηγική δρομολόγησης της πληροφορίας (unified routing) εγγυάται ότι οι μακρο-μουσικές ιδιότητες υψηλού επιπέδου, όπως η γενική δομή του κομματιού, οι ενορχηστρώσεις και οι αλλαγές των συγχορδιών, παραμένουν απόλυτα συγχρονισμένες με την κειμενική προτροπή, ακόμη και σε εκτεταμένες χρονικές κλίμακες παραγωγής.

3.4 Βελτιστοποίηση Συστήματος (Η Μηχανή Εκπαίδευσης)

Η μετάβαση μιας διατροπικής αρχιτεκτονικής βαθιάς μάθησης από ένα μη εκπαιδευμένο δίκτυο σε ένα σύστημα πρόβλεψης εξαρτάται εξ ολοκλήρου από το πλαίσιο βελτιστοποίησής της. Στη δημιουργία κειμένου σε ήχο, η μηχανή βελτιστοποίησης έχει ως στόχο να πλοηγηθεί σε ένα εξαιρετικά πολύπλοκο, πολυδιάστατο τοπίο απωλειών για να βρει μια σταθερή διαμόρφωση για εκατομμύρια ή δισεκατομμύρια παραμέτρους. Αυτή η ενότητα παρέχει μια λεπτομερή τεκμηρίωση των αλγορίθμων βελτιστοποίησης προσαρμοστικής κλίσης, των χρονοδιαγραμμάτων συνεχούς ροής και των διατυπώσεων απωλειών πολλαπλών κλιμάκων που χρησιμοποιούνται για την εκπαίδευση σύγχρονων δικτύων κειμένου σε ήχο.

3.4.1 Η Μηχανή Προσαρμοστικής Κλίσης: Adam and AdamW

Η τυπική Στοχαστική Κάθοδος Κλίσης (Stochastic Gradient Descent - SGD) κλιμακώνει τις ανανεώσεις των παραμέτρων χρησιμοποιώντας έναν ενιαίο, στατικό ρυθμό μάθησης (learning rate) που εφαρμόζεται ομοιόμορφα σε όλες τις διαστάσεις του δικτύου, βάσει του ιστορικού προτύπου των [Robbins & Monro \(1951\)](#). Αυτή η απλοποιημένη προσέγγιση συχνά αποτυγχάνει κατά την εκπαίδευση βαθιών, διατροπικών αρχιτεκτονικών που βασίζονται σε μηχανισμούς προσοχής (attention mechanisms), όπως αναλύεται διεξοδικά από τους [Goodfellow et al. \(2016\)](#).

Τα δίκτυα μετατροπής κειμένου σε ήχο (text-to-audio) καλούνται να επεξεργαστούν εξαιρετικά ασύμμετρα δεδομένα, μια ιδιαιτερότητα που επισημαίνεται στα έργα των [Liu et al. \(2023\)](#) και [Copet et al. \(2023\)](#):

- Τα αραιά διακριτικά κειμένου (sparse text tokens) παράγουν μεμονωμένα, σποραδικά σήματα κλίσεων.
- Τα πυκνά ακουστικά χαρακτηριστικά (dense audio features) δημιουργούν συνεχή, υψηλής συχνότητας σήματα κλίσεων.

Για να διασφαλιστεί η σταθερή σύγκλιση της εκπαίδευσης, τα συστήματα αυτά επιστρατεύουν αλγορίθμους προσαρμοστικού ρυθμού μάθησης (adaptive learning rate), με κυριότερους τον Adam (Adaptive Moment Estimation) και την εξελιγμένη παραλλαγή του, AdamW:

1) Adam

Στο πλαίσιο που εισήγαγαν οι [Kingma & Ba \(2014\)](#), ο αλγόριθμος Adam υπολογίζει ξεχωριστούς ρυθμούς προσαρμοστικής μάθησης για κάθε μεμονωμένη παράμετρο. Αυτό επιτυγχάνεται μέσω του υπολογισμού εκθετικά φθινόντων κινητών μέσων όρων τόσο των παρελθουσών κλίσεων (πρώτη ροπή - μοντελοποίηση ορμής) όσο και των τετραγώνων των παρελθουσών κλίσεων (δεύτερη μη κεντραρισμένη ροπή - μοντελοποίηση διακύμανσης).

Έστω θ οι εκπαιδεύσιμες παράμετροι του δικτύου και $g_t = \nabla_{\theta} L(\theta_t)$ το διάνυσμα κλίσης (gradient vector) που υπολογίζεται στο χρονικό βήμα t ως προς τη συνάρτηση απώλειας L . Η διαδικασία αυτή και ο μαθηματικός κύκλος ενημέρωσης των πρώτων και δεύτερων ροπών τυποποιούνται στη σύγχρονη βιβλιογραφία ως:

1. Ενημέρωση της Πρώτης Ροπής (Ορμή - Momentum):

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (43)$$

2. Ενημέρωση της Δεύτερης Ροπής (Κλιμάκωση Προσαρμοστικής Διακύμανσης):

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (44)$$

Όπου οι σταθερές β_1 και β_2 αποτελούν τις υπερπαραμέτρους που ελέγχουν τους εκθετικούς ρυθμούς απόσβεσης των κινητών μέσων όρων (τυπικά ορίζονται σε $\beta_1 = 0,9$ και $\beta_2 = 0,999$).

3. Διορθώσεις Αναλυτικής Μεροληψίας (Bias Correction):

Επειδή τα διανύσματα m_t και v_t αρχικοποιούνται συνήθως ως μηδενικά, εμφανίζουν μια αρχική πόλωση (bias) προς το μηδέν, ειδικά κατά τα πρώτα βήματα της εκπαίδευσης. Για την εξάλειψη αυτού του φαινομένου, εφαρμόζονται χρονικά εξαρτώμενοι παράγοντες διόρθωσης αναλυτικής μεροληψίας (bias correction):

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (45)$$

4. Βήμα Ενημέρωσης των Παραμέτρων:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \quad (46)$$

Στη σχέση αυτή, η μεταβλητή η αντιπροσωπεύει τη συνολική γραμμή βάσης του ρυθμού εκμάθησης (learning rate) και η σταθερά ϵ είναι ένας απειροελάχιστος όρος εξομάλυνσης (π.χ. 10^{-8}) που αποτρέπει την επικίνδυνη μαθηματικά διαίρεση με το μηδέν.

2) AdamW

Όταν η κανονικοποίηση L_2 (weight decay) ενσωματώνεται στον τυπικό αλγόριθμο Adam για την αποφυγή του υπερπροσαρμογής (overfitting), οι μαθηματικές πράξεις συγχωνεύονται με ανεπιθύμητο τρόπο. Ο Adam ανανεώνει τις παραμέτρους με βάση την ιστορική διακύμανση των κλίσεων (v_t). Αυτό σημαίνει ότι η ποινή της μείωσης των βαρών (weight decay penalty) κλιμακώνεται και αυτή αντιστρόφως ανάλογα με τη διακύμανση, μειώνοντας ακούσια την επίδραση της κανονικοποίησης για εκείνες τις παραμέτρους που δέχονται συχνές και μεγάλες ενημερώσεις κλίσεων. Για τη διόρθωση αυτής της δομικής ασυνέπειας, οι σύγχρονες παραγωγικές μηχανές κειμένου σε ήχο βασίζονται στο πλαίσιο AdamW, το οποίο αναπτύχθηκε από τους [Loshchilov & Hutter \(2017\)](#). Η προσέγγιση αυτή αποσυνδέει (decouples) ρητά τον υπολογισμό της φθίσης των βαρών από τις προσαρμοστικές ενημερώσεις των κλίσεων. Η ποινή κανονικοποίησης εφαρμόζεται απευθείας στο διάνυσμα των παραμέτρων αμέσως μετά την ολοκλήρωση του βήματος της κλίσης και δίνεται από τον τύπο:

$$\theta_{t+1} = \theta_t - \eta \lambda \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \quad (47)$$

Όπου η μεταβλητή λ ορίζει τον ρητό συντελεστή πτώσης των βαρών (weight decay rate).

Αυτή η αποσύνδεση διασφαλίζει ότι το μοντέλο εφαρμόζει την κανονικοποίηση ομοιόμορφα σε όλα τα επίπεδα. Με τον τρόπο αυτό, προστατεύεται η δομική και γενετική ικανότητα των επιπέδων αυτοπροσοχής πολλαπλών κεφαλών (Multi-Head Self-Attention) κατά τη διάρκεια μακροχρόνιων και απαιτητικών χρονοδιαγραμμάτων εκπαίδευσης.

3.4.2 Στόχοι Κατηγορικής και Συνεχούς Ζημίας

Ο μαθηματικός στόχος που βελτιστοποιείται από τον αλγόριθμο AdamW μεταβάλλεται ριζικά ανάλογα με τη γενετική φύση του διατροφικού συστήματος: αν δηλαδή το δίκτυο έχει

σχεδιαστεί ως ένας αυτοπαλίνδρομος προγνωστικός δείκτης διακριτών συμβόλων (tokens) ή ως ένας συνεχής μηχανισμός αποθορυβοποίησης λανθάνουσας διάχυσης (latent diffusion).

1) Κατηγορική Απώλεια Διασταυρούμενης Εντροπίας (Αυτοπαλίνδρομος Πυρήνας)

Σε συστήματα που μοντελοποιούν διαδοχικά τον ήχο αντιμετωπίζοντας τις καταχωρήσεις του βιβλίου κωδικών RVQ ως στοιχεία μιας δομημένης γλώσσας, όπως περιγράφεται από τον [Copet et al. \(2023\)](#), ο στόχος της εκπαίδευσης διαμορφώνεται ως μια εργασία κατηγορικής πρόβλεψης (categorical prediction task). Σε κάθε χρονικό βήμα, το δίκτυο καλείται να προβλέψει μια κατανομή πιθανότητας επάνω στο συνολικό προκαθορισμένο λεξιλογικό εύρος του βιβλίου κωδικών $|V|$.

Έστω y_t^j ο πραγματικός ακέραιος δείκτης (ground truth index) του ακουστικού διακριτικού στο χρονικό πλαίσιο t για το επίπεδο j της Κβαντοποίησης Υπολειμματικών Διανυσμάτων (RVQ). Αντίστοιχα, έστω $p_t^j(k)$ η τιμή πιθανότητας που υπολογίζεται από το τελικό επίπεδο Softmax του δικτύου για την καταχώρηση k του λεξιλογίου. Στο επίπεδο της αρχιτεκτονικής υλοποίησης, η πολυεπίπεδη κατηγορική απώλεια διασταυρούμενης εντροπίας (multi-codebook categorical cross-entropy loss) τυποποιείται στη σύγχρονη βιβλιογραφία ως:

$$L_{CE} = \frac{-1}{S \cdot N_q} \sum_{t=1}^S \sum_{j=1}^{N_q} \sum_{k=1}^{|V|} \mathbf{I}(y_t^j = k) \cdot \log(\hat{p}_t^j(k)) \quad (48)$$

όπου, οι μεταβλητές της αντικειμενικής συνάρτησης ορίζονται ως εξής:

- S : Το συνολικό χρονικό μήκος της ακολουθίας των πλαισίων (frames).
- N_q : Ο αριθμός των ενεργών, παράλληλων επιπέδων του βιβλίου κωδικών RVQ.
- $\mathbf{I}(\cdot)$: Η συναρτησιακή ένδειξη (indicator function), η οποία εξάγει την τιμή 1 όταν η τρέχουσα συνθήκη ταυτίζεται με τον πραγματικό δείκτη-στόχο, και 0 σε κάθε άλλη περίπτωση.

Αυτός ο τρόπος εξαναγκάζει το μοντέλο να μεγιστοποιήσει την ακριβή λογαριθμική πιθανότητα (log-likelihood) παραγωγής των σωστών χρονολογικών μουσικών διακριτικών, με βάση το καθοδηγητικό γλωσσικό πρόθεμα (text prefix) που έχει εισαχθεί.

2) Απώλεια Αποθορυβοποίησης Αντιστοίχισης Βαθμολογίας (Diffusion Core)

Για τα συστήματα που συνθέτουν ήχο μέσω της επεξεργασίας λανθάνουσας διάχυσης, όπως τεκμηριώνεται από τους [Liu et al. \(2023\)](#), το μοντέλο δεν προβαίνει στην πρόβλεψη διακριτών κατηγορικών ετικετών. Αντίθετα, στη βάση του θεωρητικού πλαισίου των [Ho et al. \(2020\)](#), το δίκτυο εκπαιδεύεται στην εκτίμηση συνεχών προφίλ θορύβου που προστίθενται προοδευτικά σε έναν πίνακα ακουστικών αναπαράστασεων υψηλής διαστατικότητας. Κατά τη φάση της εκπαίδευσης, απομονώνεται ένας αρχικός, απαλλαγμένος από αλλοιώσεις συνεχής λανθάνων τανυστής ήχου z_0 . Στη συνέχεια, δειγματοληπτείται ένα τυχαίο χρονικό βήμα προς τα εμπρός $t \in [1, T]$ και ένας τανυστής λευκού θορύβου $\epsilon \sim \mathcal{N}(0, I)$ αναμειγνύεται με το ακουστικό σήμα, διαμορφώνοντας μια αλλοιωμένη λανθάνουσα κατάσταση z_t .

Το δίκτυο αποθορυβοποίησης ϵ_θ αναλαμβάνει πλέον την υποχρέωση να προβλέψει με ακρίβεια το προφίλ του θορύβου που εισήχθη, χρησιμοποιώντας τον πίνακα κειμενικών ενσωματώσεων H_{text} για προετοιμασία. Η μαθηματική γενίκευση του απλοποιημένου στόχου απώλειας διάχυσης αποθορυβοποίησης (denoising diffusion objective) μοντελοποιείται συμβολικά ως:

$$L_{Diffusion}(\theta) = \mathbf{E}_{t, z_\theta, \epsilon} \left[\left\| \epsilon - \epsilon_\theta(z_t, t, \mathbf{H}_{text}) \right\|_2^2 \right] \quad (49)$$

Ο συγκεκριμένος στόχος εφαρμόζει έναν κλασικό υπολογισμό Μέσου Τετραγωνικού Σφάλματος (Mean Squared Error - MSE) επάνω στις συνεχείς μεταβλητές του λανθάνοντος χώρου.

Εξαναγκάζοντας τις παραμέτρους του δικτύου (θ) να ελαχιστοποιήσουν την Ευκλείδεια απόσταση ανάμεσα στον πραγματικό θόρυβο (ϵ_0) και τον εκτιμώμενο θόρυβο (ϵ_θ), ο βελτιστοποιητής χαρτογραφεί συστηματικά ένα μαθηματικό διανυσματικό πεδίο κλίσεων (score field). Κατά τον χρόνο εκτέλεσης (inference), το πεδίο αυτό καθοδηγεί με απόλυτη γεωμετρική ακρίβεια τη σταδιακή, επαναληπτική μετατροπή ενός εντελώς χαοτικού λευκού θορύβου σε ευκρινείς, δομικά συνεκτικές και καθαρές διαμορφώσεις του ακουστικού φασματογραφήματος.

3.4.3 Μοντέλο ταύτισης ροής (Flow Matching)

Ενώ τα Πιθανοτικά Μοντέλα Διάχυσης Αποθορύβωσης (Denoising Diffusion Probabilistic Models - DDPM) των [Ho et al. \(2020\)](#) βελτιστοποιούν την εκτίμηση του θορύβου πάνω σε εξαιρετικά μη γραμμικά, καμπυλωτά χρονοδιαγράμματα διακύμανσης, τα σύγχρονα χωροχρονικά δίκτυα αντικαθιστούν όλο και περισσότερο αυτούς τους τυπικούς στόχους διάχυσης. Η μετατόπιση αυτή πραγματοποιείται προς τα πλαίσια Ταύτισης Ροής (Flow Matching - FM), στη βάση του θεωρητικού υποδείγματος που εισηγήθηκε ο [Lipman \(2023\)](#). Όταν εφαρμόζεται σε λανθάνουσες ακουστικές αναπαράστασεις, η Ταύτιση Ροής παρακάμπτει πλήρως τα πολύπλοκα στοχαστικά χρονοδιαγράμματα πιθανοτήτων. Αντίθετα κατασκευάζει ένα ντετερμινιστικό, ευθύγραμμο διανυσματικό πεδίο που αντιστοιχίζει μια μη εξαρτημένη κατανομή θορύβου απευθείας στην ακουστική κατανομή-στόχο.

Μαθηματικά, έστω z_0 ένα δείγμα καθαρού Γκαουσιανού θορύβου και z_1 ένας απαλλαγμένος από αλλοιώσεις συνεχής λανθάνων πίνακας ήχου (ground truth audio). Η ευθεία διαδρομή πιθανότητας (probability path) που συνδέει αυτούς τους δύο τομείς ορίζεται μέσω μιας απλής γραμμικής παρεμβολής:

$$z_t = t z_1 + (1 - t) z_0 \quad (50)$$

Το δίκτυο παραμετροποιείται ως ένα διανυσματικό πεδίο $v_\theta(z_t, t, H_{\text{text}})$. Στόχος του είναι να προβλέψει το απόλυτο διάνυσμα της ταχύτητας-στόχου ($z_1 - z_0$) σε οποιοδήποτε ενδιάμεσο χρονικό βήμα $t \in [0, 1]$.

Η μαθηματική γενίκευση για την επίτευξη αυτού του σκοπού επιτυγχάνεται με τη βελτιστοποίηση της Απώλειας Αντιστοίχισης Ροής Υπό Συνθήκη (*Conditional Flow Matching Loss*), η οποία δίνεται ως:

$$L_{FM}(\theta) = \mathbf{E}_{t, z_0, z_t} \left[\left\| \mathbf{v}_\theta(z_t, t, \mathbf{H}_{text}) - (z_t - z_0) \right\|_2^2 \right] \quad (51)$$

Εκπαιδεύοντας τη μηχανή του AdamW ([Loshchilov & Hutter, 2017](#)) με στόχο την ελαχιστοποίηση αυτού του ευθύγραμμου μέσου τετραγωνικού σφάλματος (MSE), το τοπίο βελτιστοποίησης (optimization landscape) των παραμέτρων γίνεται σημαντικά πιο ομαλό.

Αυτή η μαθηματική ευθυγράμμιση επιταχύνει δραστικά τη σύγκλιση της εκπαίδευσης. Επιπλέον, επιτρέπει στους αλγορίθμους συμπερασμού (inference samplers) στο back-end να ανακατασκευάσουν εξαιρετικά πολύπλοκες, πολυοργανικές μουσικές συνθέσεις σε μόλις 10 έως 20 λειτουργικά βήματα. Η παραγωγή ολοκληρώνεται ταχύτατα, χωρίς να εμφανίζεται η αντιληπτική υποβάθμιση ή τα ακουστικά τεχνουργήματα (artifacts) που είναι τυπικά στα επιταχυνόμενα χρονοδιαγράμματα των μοντέλων διάχυσης.

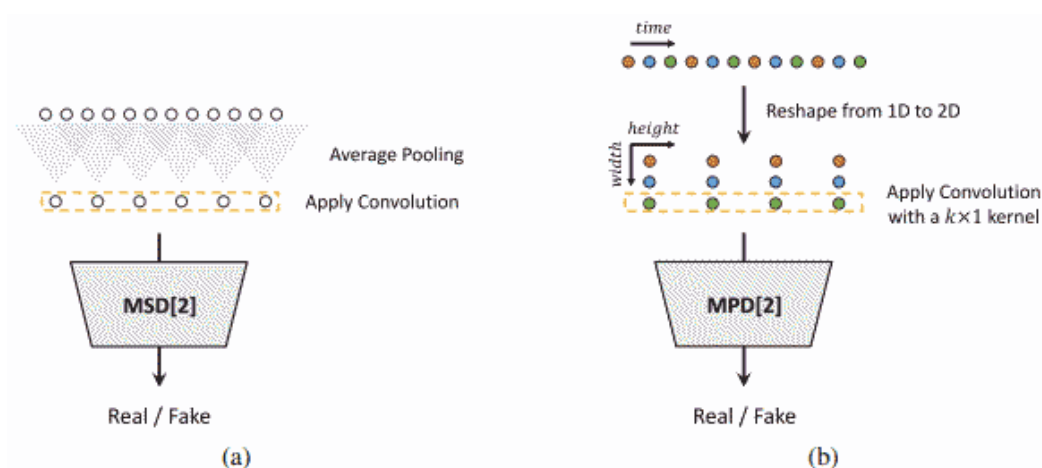
3.4.4 Αρχιτεκτονική Απώλειας Αντιστοίχισης Χαρακτηριστικών Πολλαπλών Κλιμάκων

Προκειμένου να διασφαλιστεί ότι τα συντιθέμενα μουσικά κομμάτια διατηρούν καθαρές μεταβατικές ακουστικές τιμές (transients), ακριβή στερεοφωνική κατανομή και επαγγελματικού επιπέδου ευκρίνεια στις υψηλές συχνότητες, οι σύγχρονοι αγωγοί παραγωγής απορρίπτουν τις απλοποιημένες απώλειες σε επίπεδο πλάτους (pixel-level) ή διακριτικών συμβόλων. Όταν ένα δίκτυο βελτιστοποιείται αποκλειστικά με βάση τα σφάλματα της κατηγορικής διασταυρούμενης εντροπίας ή του μέσου τετραγωνικού σφάλματος του θορύβου, οι προβλέψεις του τείνουν εγγενώς προς έναν στατιστικό μέσο όρο. Ακουστικά, αυτό το φαινόμενο εκδηλώνεται ως μια σιγασμένη έξοδος με έντονη θολότητα ως προς τη φάση (phase-blurred). Για την εισαγωγή ευκρινών και ρεαλιστικών μεταβατικών φαινομένων, η γεννήτρια πλαισιώνεται από μια στοίβα υποδικτύων που βελτιστοποιείται μέσω Απωλειών Αντιστοίχισης Χαρακτηριστικών Πολλαπλών Κλιμάκων (Multi-Scale

Feature Matching Losses), μια προσέγγιση η οποία βασίζεται στην πρωτοποριακή αρχιτεκτονική HiFi-GAN που εισήγαγαν οι [Kong et al. \(2020\)](#).

Κατά τη φάση αυτής της αντιπαραθετικής βελτιστοποίησης, η κύρια γεννήτρια κειμένου σε ήχο (G) αντιπαρατίθεται με μια συλλογή από ανεξάρτητα, εξειδικευμένα δίκτυα διακριτών (D), τα οποία χωρίζονται σε δύο κατηγορίες, η εσωτερική λειτουργική δομή των οποίων απεικονίζεται στην [Εικόνα 20](#):

- Διακριτές Φασματογράμματος Πολλαπλών Κλιμάκων (Multi-Scale Spectrogram Discriminators - MSD): Μια στοίβα από τρεις υπο-διακριτές που αξιολογούν το ακουστικό σήμα σε διαφορετικές αναλύσεις κατά μήκος του χρονικού άξονα της ακατέργαστης κυματομορφής.
- Διακριτές Κυματομορφής Πολλαπλών Περιόδων (Multi-Period Waveform Discriminators - MPD): Μια στοίβα από πέντε υπο-διακριτές, σχεδιασμένοι να απομονώνουν τις περιοδικές δομές και τις κρυφές ασυνέχειες στη φάση του ήχου.



Εικόνα 20 : Αρχιτεκτονική των δύο βασικών υποδικτύων διακριτών (sub-discriminators) του πλαισίου HiFi-GAN

Όπως φαίνεται στην [Εικόνα 20](#), η επεξεργασία του μονοδιάστατου ηχητικού σήματος διαφοροποιείται ριζικά μεταξύ των δύο αρχιτεκτονικών προσεγγίσεων:

- **Λειτουργία MSD (a):** Όπως φαίνεται στο παράδειγμα του δεύτερου υπο-διακριτή (MSD[2]), η αρχική μονοδιάστατη κυματομορφή υφίσταται διαδοχική υποδειγματοληψία μέσω επιπέδων μέσης συγκέντρωσης (Average Pooling). Η διαδικασία αυτή εξομαλύνει το σήμα σε διαφορετικές κλίμακες, επιτρέποντας στα συνελκτικά επίπεδα (Apply Convolution) να εξαγάγουν χαρακτηριστικά τόσο από τις μακροδομές (χαμηλές συχνότητες) όσο και από τις μικροδομές (υψηλές συχνότητες) του ήχου, οδηγώντας στην τελική έγκριση ή απόρριψη (Real / Fake).
- **Λειτουργία MPD (b):** Στο παράδειγμα του υπο-διακριτή MPD[2] με περίοδο $p = 3$, η μονοδιάστατη ακολουθία δειγμάτων (αποτυπώνεται με την περιοδική εναλλαγή καφέ, μπλε και πράσινων στοιχείων) αναδιατάσσεται χωρικά (Reshape from 1D to 2D). Τα στοιχεία που απέχουν κατά σταθερό χρονικό βήμα p στοιχίζονται στην ίδια στήλη ενός δισδιάστατου πίνακα. Στη συνέχεια, εφαρμόζεται μια δισδιάστατη συνέλιξη με εξειδικευμένο στενό πυρήνα διαστάσεων $k \times 1$ (Apply Convolution with a $k \times 1$ kernel). Ο σχεδιασμός αυτός επιτρέπει στον διακριτή να επεξεργάζεται ανεξάρτητα κάθε περιοδική συνιστώσα κατά μήκος του άξονα του ύψους (height), απομονώνοντας με ακρίβεια τα αρμονικά μοτίβα και τα ηχοχρώματα των μουσικών οργάνων.

Το πρωτεύον μοντέλο βελτιστοποιείται μέσω ενός συνδυασμού απωλειών πολλαπλών επιπέδων, εξασφαλίζοντας καθολική αντιληπτική ακρίβεια:

1. **Απώλεια Ανακατασκευής Φασματογράμματος Mel Πολλαπλών Κλιμάκων L_{mel} :**
Υπολογίζει την απόσταση L_I ανάμεσα στις αναπαραστάσεις φασματογράμματος log-mel της πραγματικής και της παραγόμενης κυματομορφής σε μεταβλητά μήκη παραθύρων, επιβάλλοντας ακουστική ευθυγράμμιση χαμηλού επιπέδου.
2. **Αντιπαραθετική Απώλεια (L_{adv}):** Εξαναγκάζει τη γεννήτρια να κατασκευάσει κυματομορφές ήχου που μεγιστοποιούν τις βαθμολογίες σύγχυσης (confusion scores)

των δικτύων διακριτών, κάνοντας τον συνθετικό ήχο μη ξεχωρίσιμο από τον πραγματικό.

3. **Απώλεια Αντιστοίχισης Χαρακτηριστικών Γεννήτριας (L_{FM_Adv}):** Εξάγει τις ενεργοποιήσεις των ενδιάμεσων επιπέδων (intermediate layer activations) των μπλοκ των διακριτών κατά την ανάγνωση τόσο των πραγματικών δεδομένων x όσο και των συνθετικών αρχείων (Gz).

Ο στόχος της ελαχιστοποίησης της απόστασης μεταξύ των εσωτερικών δομικών χαρακτηριστικών της γεννήτριας και των διακριτών τυποποιείται μαθηματικά ως:

$$L_{FM_Adv}(G; D) = E_{(x, z)} \left[\sum_{i=1}^M \sum_{l=1}^{L_i} \frac{1}{N_{il}} \|D_i^{(l)}(x) - D_i^{(l)}(G(z))\|_l \right] \quad (52)$$

Όπου:

- M : Ο συνολικός αριθμός των ανεξάρτητων διακριτών.
- L_i : Ο αριθμός των εσωτερικών επιπέδων (layers) στον διακριτή i .
- N_{il} : Ο συνολικός αριθμός των χαρακτηριστικών (features) στο επίπεδο l του διακριτή i .
- $\|\cdot\|_l$: Η νόρμα L_l που υπολογίζει το απόλυτο σφάλμα.

Εξαναγκάζοντας τη μηχανή βελτιστοποίησης AdamW να ελαχιστοποιήσει αυτή την πολυεπίπεδη απώλεια δομικών χαρακτηριστικών, το μοντέλο αποκτά την ικανότητα να συνθέτει εξαιρετικά λεπτές μουσικές λεπτομέρειες υψηλής πιστότητας. Η ταχεία και δυναμική έναρξη των κρουστών οργάνων (percussive onsets) ή η περίπλοκη αποσύνθεση της αντήχησης ενός χώρου (reverb decay) αποδίδονται πλέον με απόλυτο ρεαλισμό, καθώς η απώλεια αυτή αποτυπώνει χαρακτηριστικά που παραμένουν εντελώς αόρατα στις τυπικές κατηγορικές μετρήσεις.

3.5 Σύνθεση Ήχου Back-End: Ανακατασκευή της Ακουστικής Κυματομορφής κατά τον Χρόνο Εκτέλεσης

Μόλις ολοκληρωθεί η φάση της βελτιστοποίησης και οι παράμετροι του μοντέλου θ «παγώσουν», το διατροπικό σύστημα μεταβαίνει από το καθεστώς της μάθησης στη φάση της εξαγωγής συμπερασμάτων (inference) κατά τον χρόνο εκτέλεσης (runtime). Κατά τη διάρκεια του inference, οι μηχανισμοί υπολογισμού διαβαθμίσεων και ο αλγόριθμος AdamW απενεργοποιούνται πλήρως ($\nabla_{\theta} = 0$). Το νευρωνικό δίκτυο παύει να προσαρμόζει τη δομή του και συμπεριφέρεται πλέον ως μια στατική, ντετερμινιστική ή στοχαστική συνάρτηση εμπρόσθιας χαρτογράφησης (forward mapping function), σύμφωνα με τις γενικές αρχές των [Goodfellow et al. \(2016\)](#).

Ωστόσο, οι ακατέργαστες έξοδοι που παράγονται από τα επίπεδα του κεντρικού παραγωγικού πυρήνα δεν αποτελούν άμεσα αναπαραγόμενα αρχεία ήχου. Ανάλογα με την υποκείμενη αρχιτεκτονική, οι έξοδοι αυτοί εμφανίζονται ως:

- Κατηγορικά Logits Λεξιλογίου: Μη κανονικοποιημένες κατανομές πιθανοτήτων επάνω στους δείκτες του βιβλίου κωδικών (codebook) στην περίπτωση αυτοπαλίνδρομων μοντέλων.
- Συνεχείς Προβλέψεις Λανθάνοντος Θορύβου ή Ταχύτητας: Τανυστές ϵ_{θ} ή v_{θ} που αποτυπώνουν γεωμετρικές κλίσεις στην περίπτωση μοντέλων διάχυσης ή ταύτισης ροής.

Για τη μετατροπή αυτών των αφηρημένων μαθηματικών αναπαραστάσεων σε ακουστό σήμα, οι πληροφορίες διοχετεύονται σε μια εξειδικευμένη μονάδα back-end. Η μονάδα αυτή ενσωματώνει χρονοπρογραμματιστές δειγματοληψίας συμπερασμού, βρόχους αριθμητικής παρέκτασης και βαθιούς νευρωνικούς αποκωδικοποιητές, ανακατασκευάζοντας βήμα προς βήμα το τελικό ψηφιακό ηχητικό κύμα στο χρονικό πεδίο.

3.5.1 Χρονοπρογραμματιστές Δειγματοληψίας και Βρόχοι Παρέκτασης (Sampling Schedulers)

Ο τρόπος με τον οποίο το back-end διαχειρίζεται τον χρόνο εκτέλεσης εξαρτάται άμεσα από τη μαθηματική φύση του γενετικού πυρήνα:

1) Αυτοπαλίνδρομοι Αλγόριθμοι Δειγματοληψίας (Autoregressive Sampling)

Σε αυτή τη διαμόρφωση, τα logits του δικτύου μετατρέπονται σε πιθανότητες μέσω μιας συνάρτησης Softmax με ελεγχόμενη θερμοκρασία (T). Προκειμένου να αποφευχθεί η επανάληψη ή η παραγωγή ακουστικών τεχνουργημάτων (artifacts), επιστρατεύονται στοχαστικές στρατηγικές όπως η δειγματοληψία πυρήνα (Nucleus Sampling / top-p) ή η δειγματοληψία top-k, βάσει των προτύπων των [Holtzman et al. \(2020\)](#) και [Copet et al. \(2023\)](#). Ο αλγόριθμος απομονώνει ένα δυναμικό υποσύνολο του λεξιλογίου V που συγκεντρώνει το μεγαλύτερο ποσοστό της πιθανότητας, εκτελεί στοχαστική δειγματοληψία και τροφοδοτεί τον επιλεγμένο δείκτη k πίσω στο μοντέλο ως είσοδο για το επόμενο χρονικό βήμα $t+1$.

2) Αλγόριθμοι Επίλυσης Διαφορικών Εξισώσεων (Diffusion & Flow Matching Samplers)

Στα μοντέλα διάχυσης και ταύτισης ροής, ο χρονοπρογραμματιστής εκτελεί έναν επαναληπτικό βρόχο αριθμητικής επίλυσης Συνήθων Διαφορικών Εξισώσεων (ODE) ή Στοχαστικών Διαφορικών Εξισώσεων (SDE), ακολουθώντας τη θεωρητική βάση των [Ho et al. \(2020\)](#) και [Lipman et al. \(2023\)](#).

Στην περίπτωση της Ταύτισης Ροής (Flow Matching), ο δειγματολήπτης ξεκινά από έναν τυχαίο θόρυβο $z_0 \sim \mathcal{N}(0, I)$ και εκτελεί βήματα παρέκτασης κατά μήκος του προβλεπόμενου διανυσματικού πεδίου ταχύτητας v_θ , χρησιμοποιώντας επιλύτες όπως η μέθοδος Euler ή Heun. Για ένα ορισμένο μέγεθος βήματος $\Delta t = 1 / N_{steps}$, ο λανθάνων πίνακας ενημερώνεται στο επίπεδο της αρχιτεκτονικής υλοποίησης ως:

$$\mathbf{z}_{t+\Delta t} = \mathbf{z}_t + \Delta t \cdot \mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{H}_{conditioned}) \quad (53)$$

Με την ολοκλήρωση των N_{steps} βημάτων (όπου ο αριθμός N_{steps} συνήθως κυμαίνεται μεταξύ 10 και 25), ο αλγόριθμος ανακατασκευάζει πλήρως τον καθαρό, συμπιεσμένο λανθάνοντα πίνακα χαρακτηριστικών του ήχου $\mathbf{z} \in \mathbb{R}^{S \times D}$, όπως τεκμηριώνεται στο έργο των [Lipman et al. \(2023\)](#).

3.5.2 Νευρωνικοί Ακουστικοί Αποκωδικοποιητές (Neural Vocoders)

Το τελικό στάδιο της επεξεργασίας back-end αναλαμβάνει ο νευρωνικός ακουστικός αποκωδικοποιητής (neural vocoder), ο οποίος εκτελεί την αντίστροφη λειτουργία από αυτήν του κωδικοποιητή front-end. Στη σύγχρονη βιβλιογραφία, στη βάση των μοντέλων που πρότειναν οι [Zeghidour et al. \(2021\)](#) και [Kong et al. \(2020\)](#), ο αποκωδικοποιητής f_{dec} δέχεται τον ανακατασκευασμένο πίνακα $\mathbf{z}_l$ (ή τη γραμμικά αθροισμένη στοίβα RVQ $\mathbf{z}_q$) και τον μετατρέπει στην τελική κυματομορφή υψηλής διαστατικότητας:

$$\hat{\mathbf{x}} = f_{dec}(\mathbf{z}_q) \in \mathbb{R}^{T \times C} \quad (54)$$

Η αρχιτεκτονική του αποκωδικοποιητή υλοποιείται μέσω μιας σειράς από στοιβαγμένα επίπεδα ανεστραμμένης μονοδιάστατης συνέλιξης (1D transposed convolutions). Τα επίπεδα αυτά εκτελούν προοδευτική υπερδειγματοληψία (upsampling), αυξάνοντας τη συχνότητα των πλαισίων (π.χ. από τα 50 Hz του λανθάνοντος χώρου) ώστε να εναρμονιστούν με τον φυσικό ρυθμό δειγματοληψίας του ηχητικού σήματος (π.χ. 44,1 kHz). Οι συνέλιξεις υπερδειγματοληψίας συνδυάζονται με παράλληλα υπολειμματικά μπλοκ σύντηξης πολλαπλών πεδίων δεκτικότητας (Multi-Receptive Field Fusion - MRF). Τα μπλοκ αυτά έχουν εκπαιδευτεί κατά την αντιπαραθετική φάση να αποκαθιστούν με απόλυτη ακρίβεια τις σχέσεις φάσης μεταξύ των συχνοτήτων.

Με αυτόν τον τρόπο, το back-end εξαλείφει τα ψηφιακά κενά, αναδομεί τις υψηλές αρμονικές συχνότητες και εξάγει ένα συνεχές, στερεοφωνικό ακουστικό σήμα

επαγγελματικής πιστότητας, το οποίο αντιστοιχεί με ακρίβεια στις γλωσσικές απαιτήσεις της αρχικής προτροπής.

3.5.3 Στοχαστική Δειγματοληψία και Κλάδεμα Token

Στις αυτοπαλίνδρομες (autoregressive) αρχιτεκτονικές παραγωγής ήχου από κείμενο, το τελικό γραμμικό επίπεδο της τροχιάς εξάγει ένα διάνυσμα μη κανονικοποιημένων λογαριθμικών πιθανοτήτων (logits) $l_t \in \mathbb{R}^{|V|}$ για κάθε ροή του βιβλίου κωδικών (codebook) σε κάθε αυτόνομο χρονικό βήμα, όπως τεκμηριώνεται από τους [Copet et al. \(2023\)](#). Η διαχείριση αυτών των τιμών κατά τον χρόνο εκτέλεσης (inference) καθορίζει σε απόλυτο βαθμό την τελική ακουστική ποιότητα.

Η απευθείας διέλευση των logits μέσω μιας ντετερμινιστικής επιλογής μέγιστης πιθανότητας (όπως η άπληστη αποκωδικοποίηση / greedy decoding ή ο τελεστής $\arg\max$) στερεί από το μοντέλο τη δημιουργικότητα, οδηγώντας σε εξαιρετικά επαναλαμβανόμενες, στείρες και αφύσικες μουσικές συνθέσεις, σύμφωνα με την ανάλυση των [Holtzman et al. \(2020\)](#).

Αντίθετα, η μη περιορισμένη τυχαία δειγματοληψία σε ολόκληρο το εύρος του λεξιλογίου εισάγει διακριτικά στοιχεία (tokens) εξαιρετικά χαμηλής πιθανότητας, τα οποία εκδηλώνονται ακουστικά ως σοβαρά ψηφιακά τεχνουργήματα (artifacts), ακαριαία κλικ φάσης (phase clicks) και χαοτικές εκρήξεις ανεξέλεγκτων συχνοτήτων.

Προκειμένου να ενισχυθεί η μουσικότητα, η ροή και η δομική συνέπεια των κομματιών, οι σύγχρονες μηχανές συμπερασμού εφαρμόζουν στοχαστικούς αλγόριθμους κλάδεματος του λεξιλογίου, με κυριότερους τη Δειγματοληψία Top-k ([Fan et al., 2018](#)) και τη Δειγματοληψία Top-p (Πυρήνα / Nucleus Sampling) ([Holtzman et al., 2019](#)):

1) Κλιμάκωση Θερμοκρασίας (Temperature Scaling)

Πριν από την τελική επιλογή των ακουστικών διακριτικών, το σχήμα της κατανομής των logits επανακλιμακώνεται μαθηματικά με τη χρήση μιας συνεχούς, θετικής υπερπαραμέτρου θερμοκρασίας $T > 0$, στη βάση του ιστορικού προτύπου των [Ackley et al. \(1985\)](#):

$$P(t_i | t_{<t}) = \frac{\exp(l_i/T)}{\sum_{j=1}^{|V|} \exp(l_j/T)} \quad (55)$$

Η ρύθμιση της τιμής T επηρεάζει τις πιθανότητες ως εξής:

- Αν $T > 1$ (υψηλή θερμοκρασία): Ισοπεδώνει την κατανομή (flattening), εξισώνοντας τις πιθανότητες μεταξύ των διαφορετικών διακριτικών. Αυτό αυξάνει την ακουστική ποικιλομορφία και ενθαρρύνει τον μουσικό πειραματισμό, ενέχει όμως τον κίνδυνο πρόκλησης δομικής ασυνέπειας και ακουστικών σφαλμάτων.
- Αν $T \rightarrow 0$ (χαμηλή θερμοκρασία): Οξύνει την κατανομή (sharpening), μεγιστοποιώντας την απόσταση μεταξύ των επικρατέστερων επιλογών και των υπόλοιπων τιμών. Η στρατηγική αυτή αναγκάζει το μοντέλο να επιλέξει αποκλειστικά τα «ασφαλέστερα» και πιο προβλέψιμα μουσικά μοτίβα. .

2) Κλάδεμα Top-p (Πυρήνα / Nucleus Pruning)

Μετά την προσαρμογή μέσω της θερμοκρασίας, ο αλγόριθμος δειγματοληψίας πυρήνας (Nucleus Sampling) αναλαμβάνει να φιλτράρει δυναμικά τις διαθέσιμες καταχωρήσεις του λεξιλογίου, όπως προτάθηκε από τους [Holtzman et al. \(2020\)](#). Αντί να διατηρεί έναν αυστηρά σταθερό αριθμό κορυφαίων διακριτικών (όπως συμβαίνει στην κλασική μέθοδο Top-k), το Top-p απομονώνει το μικρότερο δυνατό υποσύνολο καταχωρήσεων $V^{(p)} \subset V$, των οποίων η συνδυαστική, αθροιστική μάζα πιθανότητας καλύπτει ή υπερβαίνει ένα προκαθορισμένο όριο $p \in [0, 1]$ σύμφωνα με τον τύπο:

$$\sum_{i \in V^{(p)}} P(t_i \vee t_{<t}) \geq p \quad (56)$$

Όλα τα ακουστικά διακριτικά που βρίσκονται εκτός αυτού του δυναμικού υποσυνόλου $V^{(p)}$ αποκλείονται ακαριαία: οι βαθμολογίες τους αντικαθίστανται από την τιμή $-\infty$, με

Διπλωματική Εργασία

αποτέλεσμα να αφαιρούνται πλήρως από τον πίνακα επιλογής. Στη συνέχεια, τα εναπομείναντα έγκυρα διακριτικά επανακανονικοποιούνται (re-normalized) μέσω μιας νέας συνάρτησης Softmax, διαμορφώνοντας μια καθαρή κατανομή. Από αυτή την τελική κατανομή εκτελείται μια στοχαστική κατηγορική δειγματοληψία (categorical statistical draw) για την απομόνωση του οριστικού διακριτικού ήχου(audio token) που θα κλειδώσει στο τρέχον πλαίσιο.

Αυτή η δυναμική προσαρμογή των ορίων εξασφαλίζει εξαιρετική ισορροπία. Όταν η γεννητική μηχανή είναι απόλυτα σίγουρη για την πορεία του σήματος (για παράδειγμα, κατά τη διατήρηση ενός αυστηρού και επαναλαμβανόμενου ρυθμού τυμπάνων), ο χώρος δειγματοληψίας στενεύει αυτόματα, εκμηδενίζοντας την πιθανότητα εμφάνισης ακουστικών τεχνουργημάτων. Αντίθετα, σε στιγμές ασάφειας, όπως σε μεταβάσεις κομματιών ή σύνθετες οργανικές αυτοσχεδιαστικές γέφυρες, το παράθυρο δειγματοληψίας διευρύνεται οργανικά. Αυτό επιτρέπει στο μοντέλο να ενσωματώσει δημιουργικές, απρόβλεπτες και πλούσιες μουσικές παραλλαγές, προσομοιάζοντας τη φυσική ανθρώπινη σύνθεση.

3.5.4 Ο Βρόχος Συμπερασμάτων Διπλής Διέλευσης και Καθοδήγηση Χωρίς Ταξινόμητή (CFG)

Η μαθηματική υλοποίηση της κειμενικής συνθήκης κατά τον χρόνο εκτέλεσης (inference) επιτυγχάνεται μέσω της ρητής ενσωμάτωσης του αλγορίθμου Καθοδήγησης Χωρίς Ταξινόμητή (Classifier-Free Guidance - CFG), ο οποίος προτάθηκε από τους [Ho & Salimans \(2022\)](#). Σε κάθε επαναληπτικό βήμα της ακολουθίας συμπερασμού —είτε πρόκειται για την παραγωγή των κυλιόμενων κατηγορικών logits σε μια αυτοπαλίνδρομη αλυσίδα, σύμφωνα με τους [Copet et al. \(2023\)](#), είτε για την εκτίμηση των διανυσμάτων ταχύτητας και θορύβου σε έναν πυρήνα διάχυσης, όπως τεκμηριώνεται στα έργα των [Ho et al. \(2020\)](#) και [Liu et al. \(2023\)](#)— το μοντέλο δεν εκτελεί μία μόνο διέλευση προς τα εμπρός (forward pass). Αντίθετα, πραγματοποιεί έναν υπολογισμό διπλής διέλευσης προς τα εμπρός (double forward pass).

Για ένα σύστημα λανθάνουσας διάχυσης στο τρέχον βήμα συμπερασμού t , το γλωσσικό front-end μεταβιβάζει παράλληλα στο δίκτυο δύο διαφορετικές οντότητες: τον σαφή πίνακα συνθήκης του χρήστη H_{text} και το εκπαιδευμένο, στατικό διάνυσμα κενού κειμένου (null-prompt embedding) $H_{\emptyset}$. Ο πυρήνας αποθορυβοποίησης υπολογίζει στη συνέχεια δύο ανεξάρτητες εκτιμήσεις θορύβου:

- Υπό συνθήκη εκτίμηση (Conditioned): $\epsilon_{\theta}(z_t, t, H_{text})$
- Άνευ όρων εκτίμηση (Unconditioned): $\epsilon_{\theta}(z_t, t, H_{\emptyset})$

Στη σύγχρονη βιβλιογραφία, η μαθηματική γενίκευση για την κατασκευή του τελικού διανύσματος ϵ_t το οποίο θα καθοδηγήσει τον χρονοπρογραμματιστή για την αφαίρεση του θορύβου, υλοποιείται μέσω γραμμικής παρέκτασης (linear extrapolation) μακριά από την άνευ όρων γραμμή βάσης, με τη χρήση μιας βαθμωτής παραμέτρου κλίμακας καθοδήγησης (guidance scale) $s \geq 1$, δίνεται από τον τύπο:

$$\hat{\epsilon}_t = \epsilon_{\theta}(z_t, t, \mathbf{H}_{\emptyset}) + s \cdot [\epsilon_{\theta}(z_t, t, \mathbf{H}_{text}) - \epsilon_{\theta}(z_t, t, \mathbf{H}_{\emptyset})] \quad (57)$$

Με τη δυναμική ρύθμιση του βαθμωτού συντελεστή s κατά το inference, ο χρήστης αποκτά άμεσο έλεγχο επάνω στη γενετική συμπεριφορά και την ακουστική πιστότητα της σύνθεσης:

- Αν $1 < s < 4$ (τυπική κλίμακα) : Εξασφαλίζει βέλτιστη ισορροπία ανάμεσα στον ακουστικό πλουραλισμό (diversity) και τη βασική συμμόρφωση προς τις κειμενικές οδηγίες της προτροπής.
- Αν $s > 7$ (υψηλή κλίμακα) : Ενισχύει επιθετικά τα μαθηματικά και σημασιολογικά χαρακτηριστικά που είναι μοναδικά για την προτροπή της εξαρτημένης επεξεργασίας, ενώ ταυτόχρονα καταστέλλει βίαια κάθε άσχετη ακουστική παραλλαγή. Η προσέγγιση αυτή επιβάλλει μια εξαιρετικά αυστηρή προσήλωση στο κείμενο.

Η επιθετική ενίσχυση μέσω πολύ υψηλών τιμών του s περιορίζει δραστικά τον ελεύθερο χώρο δημιουργικής κατανομής του μοντέλου. Αυτή η γεωμετρική και στατιστική συμπίεση των λανθανουσών τιμών ενδέχεται περιστασιακά να εκδηλωθεί ως ακουστικός κορεσμός, ψηφιακή παραμόρφωση (saturation artifacts) ή ψαλιδισμός (clipping) στην τελική παραγόμενη κυματομορφή.

3.5.5 Χρονοπρογραμματιστές για Προφίλ Συνεχούς Διάχυσης: Ντετερμινιστική Αποθορυβοποίηση

Στα αρχιτεκτονικά πλαίσια λανθάνουσας διάχυσης κειμένου σε ήχο (text-to-audio latent diffusion), η παραγωγή του ακουστικού σήματος επιτυγχάνεται μέσω της εκτέλεσης μιας αντίστροφης μαθηματικής διαδρομής, όπως τεκμηριώνεται από τους [Liu et al. \(2023\)](#). Η διαδικασία εκκινεί με έναν τανυστή καθαρού λευκού θορύβου Gauss $z_T \sim \mathcal{N}(0, I)$. Στη συνέχεια, επιστρατεύεται το εξαρτημένο από το κείμενο δίκτυο αποθορυβοποίησης ϵ_θ , προκειμένου να αφαιρεθούν επαναληπτικά συγκεκριμένα κλάσματα θορύβου κατά μήκος μιας προκαθορισμένης ακολουθίας βημάτων συμπερασμού (inference steps).

Το αρχικό θεωρητικό πλαίσιο των Πιθανοτικών Μοντέλων Διάχυσης Αποθορυβοποίησης (Denoising Diffusion Probabilistic Models - DDPM) απαιτεί η τροχιά του συμπερασμού να αντικατοπτρίζει πιστά και βήμα προς βήμα την εμπρόσθια διαδρομή αλλοίωσης του σήματος. Αυτό σημαίνει ότι ο αλγόριθμος πρέπει να εκτελέσει διαδοχικά $T = 1000$ διακριτά βήματα ενημέρωσης, σύμφωνα με τη βασική μεθοδολογία των [Ho et al. \(2020\)](#). Επειδή κάθε μεμονωμένο βήμα προϋποθέτει μια πλήρη διέλευση προς τα εμπρός (forward pass) μέσα από ένα πολυεπίπεδο δίκτυο U-Net ή έναν Μετασχηματιστή Διάχυσης (Diffusion Transformer), η δημιουργία ενός και μόνο μουσικού κομματιού μέσω DDPM απαιτεί αρκετά λεπτά εντατικής επεξεργασίας στην GPU. Αυτή η υπολογιστική συμφόρηση καθιστά απαγορευτική την ανάπτυξη τέτοιων συστημάτων σε εφαρμογές πραγματικού χρόνου (real-time applications).

Για την άρση αυτού του φράγματος καθυστέρησης (latency barrier), οι σύγχρονοι διαχυτές ήχου ενσωματώνουν προηγμένους χρονοπρογραμματιστές που βασίζονται στα Έμμεσα Μοντέλα Διάχυσης Αποθορυβοποίησης (Denoising Diffusion Implicit Models - DDIM), τα

οποία εισήχθησαν από τους [Song et al. \(2022\)](#). Το πλαίσιο DDIM αναδιατυπώνει την αντίστροφη γενετική πορεία ως μια ντετερμινιστική μη Μαρκοβιανή διαδρομή απόφασης. Η δομική αυτή προσαρμογή επιτρέπει στη μηχανή συμπερασμού να παρακάμπτει μεγάλα τμήματα της χρονογραμμής, εκτελώντας δειγματοληψία επάνω σε μια επιταχυνόμενη υποακολουθία βημάτων. Με τον τρόπο αυτό, ο βρόχος συμπιέζεται δραστικά (για παράδειγμα, από τα 1000 βήματα στα 50) χωρίς να υποβαθμίζεται η τελική ακουστική πιστότητα.

Το βήμα της ντετερμινιστικής ενημέρωσης για έναν συμπιεσμένο δείκτη συμπερασμού DDIM t διατυπώνεται μαθηματικά και τυποποιείται στη σύγχρονη βιβλιογραφία μέσω της ακόλουθης εξίσωσης:

$$\mathbf{z}_{t-l} = \sqrt{\alpha_{t-l}} \left(\frac{\mathbf{z}_t - \sqrt{1 - \alpha_t} \cdot \epsilon_\theta(\mathbf{z}_t, t, \mathbf{H}_{text})}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-l} - \sigma_t^2} \cdot \epsilon_\theta(\mathbf{z}_t, t, \mathbf{H}_{text}) + \sigma_t \epsilon \quad (58)$$

Όπου:

- α_t : Ορίζει το προκαθορισμένο χρονοδιάγραμμα της ντετερμινιστικής διακύμανσης (variance schedule) στο χρονικό βήμα t .
- ϵ_θ : Αντιπροσωπεύει το προφίλ του θορύβου που εκτιμάται δυναμικά από το νευρωνικό δίκτυο, υπό την καθοδήγηση της κειμενικής ενσωμάτωσης H_{text} , όπως αναλύεται από τους [Liu et al. \(2023\)](#).
- σ_t : Αποτελεί την παράμετρο που ελέγχει τον βαθμό της στοχαστικής μεταβλητότητας (stochastic volatility) κατά την αναδρομή, σύμφωνα με τους [Song et al. \(2022\)](#).

Θέτοντας την παράμετρο στοχαστικότητας αυστηρά στο $\sigma_t = 0$, ο όρος του τυχαίου θορύβου $\sigma_t \cdot \epsilon$ εκμηδενίζεται και η γενετική διαδρομή μετατρέπεται σε μια πλήρως ντετερμινιστική τροχιά. Με βάση έναν συγκεκριμένο αρχικό πίνακα θορύβου $\mathbf{z}_T$ και μια σταθερή γλωσσική προτροπή H_{text} , ο χρονοπρογραμματιστής DDIM χαράσσει μια απόλυτα συνεπή και επαναλήψιμη διανυσματική πορεία (stable map) στον λανθάνοντα χώρο.

Μέσα σε μερικές δεκάδες επαναλήψεις, ο μηχανισμός αφαιρεί συστηματικά τα χαοτικά στοιχεία του θορύβου, ανασυνθέτοντας τις λανθάνουσες γεωμετρικές δομές. Η διαδικασία ολοκληρώνεται με τον σχηματισμό ενός εξαιρετικά δομημένου και απαλλαγμένου από αλλοιώσεις Λανθάνοντος Φασματογραφήματος Mel (Latent Mel Spectrogram), το οποίο είναι πλέον έτοιμο να διοχετευθεί στον νευρωνικό αποκωδικοποιητή για την παραγωγή του τελικού ηχητικού κύματος.

3.5.6 Επιτάχυνση Τροχιάς Υψηλής Τάξης: DPM-Solver++

Ενώ η ντετερμινιστική διατύπωση του χρονοπρογραμματιστή DDIM σπάει με επιτυχία τον περιορισμό των 1000 βημάτων των τυπικών Μαρκοβιανών αλυσίδων, συμπεριφέρεται δομικά ως ένας αριθμητικός επιλυτής Συνήθων Διαφορικών Εξισώσεων (ΣΔΕ) πρώτης τάξης, όπως περιγράφεται από τους [Song et al. \(2022\)](#). Αυτή η ιδιότητα τον καθιστά ευάλωτο σε σφάλματα περικοπής (truncation errors) στην περίπτωση που το μέγεθος του βήματος διευρυνθεί υπερβολικά —δηλαδή όταν μειωθεί δραστικά ο αριθμός των συνολικών βημάτων δειγματοληψίας. Για την επίτευξη πραγματικά real-time ρυθμών σύνθεσης ήχου χωρίς την πρόκληση παραμορφώσεων φάσης (phase distortion) ή θολώματος του ηχοχρώματος (timbral blurring), οι σύγχρονοι συνεχείς πυρήνες ήχου αντικαθιστούν την παρακολούθηση πρώτης τάξης με το εξελιγμένο πλαίσιο DPM-Solver++ ([Lu et al., 2022](#)).

Το DPM-Solver++ αναδιατυπώνει την αντίστροφη τροχιά αποθορυβοποίησης ως μια ημιγραμμική συνήθη διαφορική εξίσωση (semi-linear ODE). Αντί να υπολογίζεται το επόμενο λανθάνον βήμα z_{t+1} χρησιμοποιώντας αποκλειστικά την πρόβλεψη της στιγμιαίας κλίσης στο χρονικό βήμα t_i , ένας επιλυτής πολωνύμων υψηλής τάξης (δεύτερης ή τρίτης τάξης) ενσωματώνει το διανυσματικό πεδίο της τροχιάς λαμβάνοντας υπόψη πολλά ιστορικά βήματα ταυτόχρονα.

Στο επίπεδο της αρχιτεκτονικής υλοποίησης, η μαθηματική αντιστοίχιση της ενημέρωσης δεύτερης τάξης (2nd-order update) τυποποιείται στη σύγχρονη βιβλιογραφία ως εξής:

$$\mathbf{z}_{t_{i+1}} = \frac{\sigma_{t_{i+1}}}{\sigma_{t_i}} \mathbf{z}_{t_i} - \alpha_{t_{i+1}} (e^{h_i} - 1) \hat{\epsilon}(\mathbf{z}_{t_i}) - \frac{\alpha_{t_{i+1}} (e^{h_i} - 1)}{2 r_i} [\hat{\epsilon}(\mathbf{z}_{t_i}) - \hat{\epsilon}(\mathbf{z}_{t_{i-1}})] \quad (59)$$

Οι μεταβλητές και οι παραμέτροι της εξίσωσης ορίζονται ως εξής:

- σ_t, α_t : Αντιπροσωπεύουν τις λειτουργικές παραμέτρους του χρονοδιαγράμματος θορύβου, οι οποίες καθορίζονται με βάση τον χώρο του λογαριθμικού λόγου σήματος προς θόρυβο (log-SNR domain).
- $h_i = \lambda_{t_{i+1}} - \lambda_{t_i}$: Ορίζει το ακριβές μέγεθος του βήματος (step size) εντός του πεδίου log-SNR, όπου $\lambda = \ln(\alpha_t / \sigma_t)$.
- $r_i = \frac{\lambda_{t_{i+1}} - \lambda_{t_i}}{\lambda_{t_i} - \lambda_{t_{i-1}}}$: Αντιπροσωπεύει τον λόγο των βημάτων (step ratio) ανάμεσα στις διαδοχικές χρονικές στιγμές.
- $\hat{\epsilon}(\mathbf{z}_{t_i}), \hat{\epsilon}(\mathbf{z}_{t_{i-1}})$: Οι εκτιμήσεις του θορύβου που παράγονται από το νευρωνικό δίκτυο για το τρέχον και το αμέσως προηγούμενο χρονικό βήμα αντίστοιχα.

Επειδή το DPM-Solver++ λαμβάνει υπόψη του τη μαθηματική καμπυλότητα της τροχιάς αποθορυβοποίησης και δεν την προσεγγίζει ως μια σειρά από απλές ευθείες γραμμές, επιτυγχάνει εξαιρετική γεωμετρική σταθερότητα. Η γεωμετρική αυτή ακρίβεια επιτρέπει στο back-end του σήματος να διατηρεί ανέπαφη την ακουστική πιστότητα, προστατεύοντας τις ακαριαίες χρονικές ενάρξεις των κρουστών οργάνων (crisp percussive onsets) και τις ευαίσθητες αρμονικές ευθυγραμμίσεις των παραγόμενων συχνοτήτων. Το σημαντικότερο πλεονέκτημα είναι ότι η συγκεκριμένη ποιότητα επιτυγχάνεται με τη συμπίεση του υπολογιστικού φόρτου εργασίας στο inference στο ελάχιστο δυνατό όριο των 15 έως 25 συνολικών περασμάτων προς τα εμπρός (forward passes). Η παραγωγική διαδικασία

καθίσταται πλέον εξαιρετικά αποδοτική, επιτρέποντας τη χρήση των μοντέλων διάχυσης σε απαιτητικά περιβάλλοντα αλληλεπίδρασης και σύνθεσης σε πραγματικό χρόνο.

3.5.7 Η Τελική Αποσυμπίεση: Σύνθεση Κυματομορφής μέσω Νευρωνικών Vocoder

Στο απόλυτο σημείο εξόδου του αγωγού συμπερασμού (inference pipeline), οι παραγόμενες μεταβλητές εξακολουθούν να βρίσκονται εγκλωβισμένες μέσα σε έναν «σιωπηλό», συμπίεσμένο λανθάνοντα χώρο μαθηματικών αναπαραστάσεων. Για να μετατραπούν αυτά τα αφηρημένα χαρακτηριστικά σε ένα φυσικό, αναπαραγωγίμο ψηφιακό αρχείο ήχου στο χρονικό πεδίο (όπως μια κυματομορφή wav), πρέπει υποχρεωτικά να διέλθουν από το τελικό επίπεδο αποσυμπίεσης του back-end.

Ανάλογα με τη γενετική φύση της υποκείμενης αρχιτεκτονικού πυρήνα, η αποσυμπίεση και η τελική σύνθεση της κυματομορφής εκτελούνται μέσω δύο διακριτών τεχνολογικών διαδρομών:

1) Η Διαδρομή Αποκωδικοποίησης του Νευρωνικού Κωδικοποιητή (Neural Audio Decoder)

Σε αυτοπαλίνδρομα μοντέλα που βασίζονται στη διαπλοκή διακριτών συμβόλων (interleaved tokens), όπως το MusicGen, οι παραγόμενες ροές ακέραιων δεικτών δρομολογούνται πίσω στο μπλοκ του αποκωδικοποιητή f_{dec} του νευρωνικού κωδικοποιητή ήχου (όπως το EnCodec), σύμφωνα με το πλαίσιο των [Défossez et al. \(2022\)](#), η συνολική ροή του οποίου έχει αποτυπωθεί στην [Εικόνα 15](#) της υποενότητας 3.2.3.

Ο αποκωδικοποιητής εφαρμόζει μια αυστηρή αλυσίδα προοδευτικής υπερδειγματοληψίας (upsampling) η οποία αποτελείται από στοιβαγμένα επίπεδα μονοδιάστατων μετατεθειμένων συνελιζέων (1D transposed convolutions), διασκορπισμένα με πολυ-κλιμακωτά δίκτυα υπολοίπων (multi-scale residual networks). Η διαδικασία αυτή αντιστρέφει πλήρως την αρχική γεωμετρική χαρτογράφηση της συμπίεσης. Προβάλλει τους κωδικούς χαμηλής συχνότητας των πλαισίων (frames) πίσω στον πλήρη φυσικό ρυθμό δειγματοληψίας των 44

kHz, αποκαθιστώντας τις ομαλές ευθυγραμμίσεις φάσης κατά μήκος ολόκληρης της συνεχούς ακουστικής χρονικής γραμμής.

2) Η Διαδρομή του Νευρωνικού Vocoder (Neural Vocoder Path)

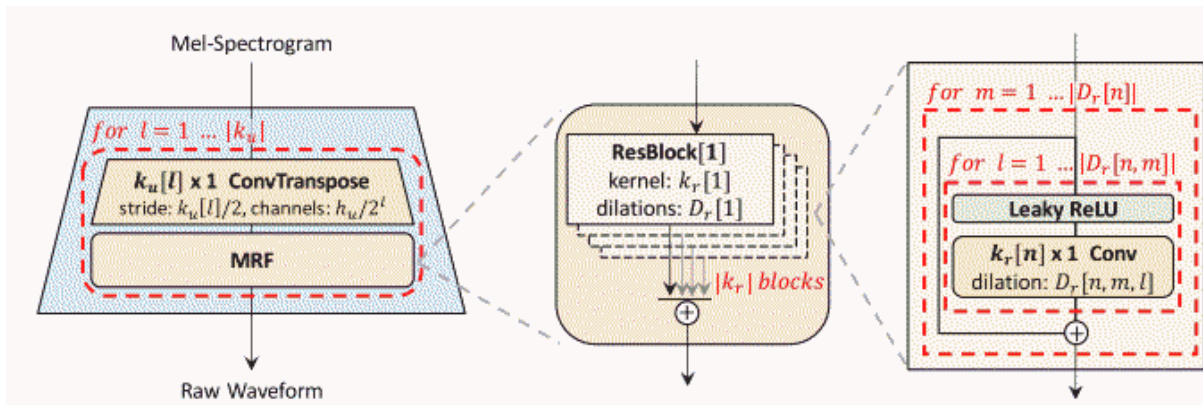
Σε παραγωγικά συστήματα διάχυσης (diffusion) ή ταύτισης ροής (flow matching) που ειδικεύονται στη δημιουργία φασματογραφημάτων, όπως το AudioLDM που περιγράφεται από τους [Liu et al. \(2023\)](#), οι έξοδοι εξέρχονται από τους επιταχυνόμενους επιλύτες (DDIM ή DPM-Solver++) με τη μορφή μιας δισδιάστατης εικόνας (2D matrix) που αναπαριστά ένα Log-Mel Φασματογράφημα.

Επειδή το φασματογράφημα Mel αποτελεί μια αναπαράσταση μόνο μέτρου/μεγέθους (magnitude-only matrix) από την οποία απουσιάζει η πληροφορία της φάσης, η απευθείας μετατροπή του σε ήχο είναι αδύνατη. Για τη γεφύρωση αυτού του χάσματος και τη χαρτογράφηση του πίνακα πίσω σε μια κυματομορφή χρονικού πεδίου, ο αγωγός επιστρατεύει ένα αυτόνομο, βαθύ δίκτυο νευρωνικού κωδικοποιητή φωνής (neural vocoder), με επικρατέστερη επιλογή την αρχιτεκτονική HiFi-GAN ([Kong et al., 2020](#)) της οποίας η γεννήτρια εκτελεί μια ιεραρχική αποδόμηση και ανασύνθεση του σήματος αποτυπώνεται και στην [Εικόνα 21](#).

- **Υπερδειγματοληψία (Left Module):** Το Log-Mel Φασματογράφημα εισέρχεται στο δίκτυο και περνά από μια σειρά $|k_u|$ διαδοχικών επιπέδων αναστραμμένης συνέλιξης (1D ConvTranspose). Κάθε επίπεδο l χρησιμοποιεί έναν εξειδικευμένο πυρήνα $k_u[l]$ και ένα βήμα (stride) $k_u[l]/2$ αυξάνοντας προοδευτικά τη χρονική ανάλυση του σήματος ώστε να προσεγγίσει τον ρυθμό της ακατέργαστης κυματομορφής (Raw Waveform).
- **Μονάδα Σύντηξης Πολλαπλών Πεδίων Δεκτικότητας (MRF - Middle Module):** Μετά από κάθε βήμα υπερδειγματοληψίας, το σήμα διοχετεύεται παράλληλα σε μια μονάδα MRF (Multi-Receptive Field Fusion). Όπως φαίνεται στο διάγραμμα, το MRF αποτελείται από μια στοίβα $|k_r|$ παράλληλων υπολειμματικών μπλοκ (ResBlock[1],

ResBlock[2], κλπ.), καθένα από τα οποία διαθέτει διαφορετικό μέγεθος πυρήνα $k_r[l]$ και διαφορετικούς ρυθμούς διεύρυνσης (dilations) $D_r[l]$. Οι έξοδοι αυτών των παράλληλων δικτύων αθροίζονται, επιτρέποντας στο μοντέλο να συνθέτει ταυτόχρονα χαρακτηριστικά από ποικίλα αντιληπτικά πεδία.

- **Μικρο-Δομή του Υπολειμματικού Μπλοκ (Right Module):** Στο δεξί τμήμα αποκαλύπτεται η εσωτερική ροή του n -οστού ResBlock. Η πληροφορία περνά επαναληπτικά για $m = 1 \dots |D_r[l]|$ επίπεδα, όπου κάθε επίπεδο εφαρμόζει μια μη γραμμικότητα Leaky ReLU ακολουθούμενη από μια μονοδιάστατη συνέλιξη με εκθετικά αυξανόμενο ρυθμό διεύρυνσης ($D_r[n, m, l]$). Αυτός ο σχεδιασμός είναι που επιτρέπει στο HiFi-GAN να επιλύει το πρόβλημα της φάσης, εξασφαλίζοντας κορυφαία υπολογιστική ταχύτητα και εξαλείφοντας τα μεταλλικά τεχνουργήματα.



Εικόνα 21: Αρχιτεκτονική της γεννήτριας (generator) του HiFi-GAN

Μέσα από αυτή τη διαδικασία, ο vocoder καταφέρει να προβλέψει ταυτόχρονα και με απόλυτη συγχρονικότητα τις υποκείμενες περιόδους του θεμελιώδους ύψους/συχνότητας (pitch periods), τις περίπλοκες αρμονικές υπερθέσεις (harmonic overtones) των οργάνων και τις ακριβείς ευθυγραμμίσεις φάσης (phase alignments) μεταξύ των συχνотήτων.

Συνδυάζοντας τους προηγμένους χρονοπρογραμματιστές δειγματοληψίας με τους νευρωνικούς vocoders υψηλής πιστότητας, η μονάδα σύνθεσης ήχου του back-end ολοκληρώνει επιτυχώς τον διατροφικό κύκλο ζωής του συστήματος. Μετατρέπει τις αρχικές, αφηρημένες γλωσσικές ενσωματώσεις του κειμένου, τους δομημένους χάρτες

διασταυρούμενης προσοχής και τις λανθάνουσες μαθηματικές κατανομές σε μια συνεχή, φυσική ροή ακουστικών σημάτων υψηλής ακουστικής πιστότητας. Το τελικό αποτέλεσμα είναι απόλυτα εναρμονισμένο —τόσο δομικά όσο και συναισθηματικά— με την πρωταρχική ανθρώπινη προτροπή.

4. Συγκριτική Ανάλυση Γενετικών Παραδειγμάτων

4.1 Αυτοπαλίνδρομη Μοντελοποίηση Γλώσσας και Διαδοχική Παραγωγή Ακουστικών Διακριτικών

Το παράδειγμα της Αυτοπαλίνδρομης (Autoregressive - AR) Μοντελοποίησης Γλώσσας αντιπροσωπεύει μια βαθιά εννοιολογική μετατόπιση στο πεδίο της πολυτροπικής σύνθεσης ήχου. Η προσέγγιση αυτή δεν αντιμετωπίζει τη δημιουργία μουσικής ως ένα κλασικό πρόβλημα ακουστικής μηχανικής ή επεξεργασίας σημάτων, αλλά ως μια συμβολική εργασία παραγωγής ακολουθίας (sequence generation task), όπως τεκμηριώνεται στα έργα των [Vaswani et al. \(2017\)](#) και [Borsos et al. \(2023\)](#). Μόλις ο νευρωνικός κωδικοποιητής ήχου προβάλει και διακριτοποιήσει τις ακατέργαστες συνεχείς κυματομορφές σε πολυεπίπεδα πλέγματα διακριτών δεικτών (codebook indices), η μουσική μεταφράζεται αποτελεσματικά σε μια εξαιρετικά δομημένη «γλώσσα» ([Zeghidour et al., 2021](#); [Défossez et al., 2022](#)).

Κατά συνέπεια, οι σύγχρονες αυτοπαλίνδρομες αρχιτεκτονικές —με κυριότερα παραδείγματα το πλαίσιο MusicGen / AudioCraft ([Copet et al. \(2023\)](#)) και τον δομικό προκάτοχό του AudioLM ([Borsos et al., 2023](#))— αξιοποιούν βαθιές δομές Transformer που αποτελούνται αποκλειστικά από αποκωδικοποιητές (decoder-only). Στόχος τους είναι η διαδοχική μοντελοποίηση της από κοινού κατανομής πιθανότητας αυτών των ακουστικών διακριτικών (audio tokens).

4.1.1 Μαθηματικό Πλαίσιο και Εξάρτηση Ακολουθίας

Μαθηματικά, έστω ότι ο πίνακας $U \in \mathbb{N}^{S \times N_q}$ αντιπροσωπεύει ένα πλέγμα διακριτών ακουστικών tokens που εξάγονται από έναν αγωγό Κβαντοποίησης Υπολειμματικών Διανυσμάτων (RVQ). Στη δομή αυτή, η μεταβλητή S ορίζει το συνολικό χρονικό μήκος της ακολουθίας (συνολικά πλαίσια / frames) και η μεταβλητή N_q αντιπροσωπεύει τον αριθμό των παράλληλων επιπέδων του βιβλίου κωδικών.

Ο θεμελιώδης στόχος του αυτοπαλίνδρομου μοντέλου είναι να παραγοντοποιήσει την από κοινού πιθανότητα της ακολουθίας των ακουστικών δεικτών σε μια αλυσίδα υπό όρους πιθανοτήτων, εφαρμόζοντας τον κλασικό αλυσιδωτό κανόνα των πιθανοτήτων (chain rule of probability):

$$P(U \vee H_{text}) = \prod_{t=1}^S P(u_t \vee u_{<t}, H_{text}) \quad (60)$$

Όπου:

- $u_t = [u_t^1, u_t^2, \dots, u_t^{N_q}]$: Αντιπροσωπεύει το διάνυσμα των παράλληλων διακριτικών όλων των επιπέδων του codebook στο συγκεκριμένο χρονικό πλαίσιο t .
- $u_{<t}$: Υποδηλώνει το πλήρες ιστορικό των παραχθέντων ακουστικών tokens από την έναρξη της αλληλουχίας μέχρι τη χρονική στιγμή $t-1$.
- H_{text} : Αντιπροσωπεύει τον πυκνό σημασιολογικό πίνακα που εξήχθη από τον μηχανισμό επεξεργασίας και κωδικοποίησης των κειμενικών προτροπών (text front-end).

Επειδή η αρχιτεκτονική RVQ κατασκευάζει ένα παράλληλο πλέγμα πολλαπλών επιπέδων από διακριτά tokens αντί για μία γραμμική, μονοδιάστατη ροή, ένα τυπικό γλωσσικό μοντέλο αδυνατεί να επεξεργαστεί τα δεδομένα απευθείας χωρίς ριζικές δομικές τροποποιήσεις. Εάν το σύστημα επιχειρήσει να προβλέψει ταυτόχρονα όλους τους δείκτες των N_q βιβλίων κωδικών για το πλαίσιο t , θα απαιτούνταν πολλαπλές ανεξάρτητες κεφαλές εξόδου (prediction heads). Οι κεφαλές αυτές, όμως, θα λειτουργούσαν «τυφλά» ως προς τις ενεργές επιλογές των διπλανών επιπέδων κατά το ίδιο χρονικό βήμα. Αυτή η αποτυχία συντονισμού (coordination failure) οδηγεί σε σοβαρές κάθετες ακουστικές αντιφάσεις και ασυνέπειες (για παράδειγμα, ένα διακριτικό που αντιστοιχεί σε κρούση τυμπάνου στο επίπεδο 1 να συγκρούεται βίαια με μια φωνητική υφή που προβλέπεται ταυτόχρονα στο επίπεδο 2, καταστρέφοντας το ηχόχρωμα).

Για την άρση αυτού του αδιεξόδου, οι σύγχρονες αυτοπαλίνδρομες μηχανές εισάγουν εξειδικευμένα Μοτίβα Διαπλοκής και Καθυστέρησης Βιβλίου Κωδικών (Codebook Interleaving and Delay Patterns), όπως αναδεικνύεται στο έργο των [Copet et al. \(2023\)](#). Η πιο διαδεδομένη και αποδοτική διαμόρφωση είναι το Μοτίβο Καθυστέρησης (Delay Pattern), κατά το οποίο εισάγεται μια τεχνητή, σταθερή χρονολογική μετατόπιση ανάμεσα στα παράλληλα επίπεδα του codebook. Μετατοπίζοντας κάθε επόμενο επίπεδο του βιβλίου κωδικών προς τα εμπρός κατά ακριβώς ένα χρονικό βήμα, ο αρχικός πίνακας των πολλαπλών παράλληλων ροών ισοπεδώνεται (flattened) και μετατρέπεται σε μια ενιαία, αιτιακά ξετυλιγμένη ακολουθία (causally unrolled sequence).

Όταν ο Μετασχηματιστής (Transformer) επεξεργάζεται το τρέχον πλαίσιο, ο εσωτερικός πίνακας της αιτιακής αυτοπροσοχής (causal self-attention matrix) επιτρέπει στο επίπεδο 2 του βιβλίου κωδικών να «κοιτάζει» ενεργά προς τα πίσω, σύμφωνα με τις προδιαγραφές του [Vaswani \(2017\)](#). Με αυτόν τον τρόπο, εξαρτά και προσαρμόζει μαθηματικά την πρόβλεψη του από την επιλογή που πραγματοποίησε το επίπεδο 1 στο αμέσως προηγούμενο υπολογιστικό βήμα. Αυτή η κομψή διαπλοκή επιτρέπει σε έναν κλασικό αιτιακό πυρήνα Transformer να παράγει εξαιρετικά συγχρονισμένα, πολυεπίπεδα και αρμονικά συνεκτικά κομμάτια ορχηστρικών οργάνων, εκτελώντας ολόκληρη τη διαδικασία μέσα σε ένα και μόνο υπολογιστικό πέρασμα προς τα εμπρός (single forward pass).

4.1.2 Αναλυτική αξιολόγηση: Πλεονεκτήματα και Μειονεκτήματα

Η ευρεία υιοθέτηση της αυτοπαλίνδρομης (Autoregressive - AR) μοντελοποίησης σε πρωτοποριακές μηχανές σύνθεσης μουσικής, όπως το Jukebox των [Dhariwal et al. \(2020\)](#) και το MusicGen των [Copet et al. \(2023\)](#), υπογραμμίζει την εξαιρετική ικανότητα της αρχιτεκτονικής αυτής να διαχειρίζεται αυστηρές δομικές ιεραρχίες. Παράλληλα, όμως, αναδεικνύει και σημαντικούς περιορισμούς που επηρεάζουν καταλυτικά την υπολογιστική της αποδοτικότητα κατά τον χρόνο εκτέλεσης.

1) Πλεονεκτήματα

- Άριστη Μακροδομική και Χρονολογική Συνοχή

Η μουσική διέπεται από αυστηρές χρονολογικές εξαρτήσεις μεγάλης εμβέλειας (long-range dependencies), οι οποίες περιλαμβάνουν επαναλαμβανόμενα ρυθμικά μοτίβα (time signatures), διακριτές δομικές ενότητες (όπως η εναλλαγή κουπλέ-ρεφρέν) και απόλυτη ρυθμική ακρίβεια. Επειδή τα αυτοπαλίνδρομα μοντέλα χρησιμοποιούν μια αυστηρή μάσκα αιτιακής προσοχής (causal attention mask), κάθε διακριτικό στοιχείο (token) που δημιουργείται στο χρονικό βήμα t αναγκάζεται να κοιτάζει προς τα πίσω, σύμφωνα με το κλασικό μοντέλο των [Vaswani et al. \(2017\)](#). Υπολογίζει έτσι ρητές εξαρτήσεις εσωτερικού γινομένου (dot-product attention) επάνω σε ολόκληρο το χρονολογικό ιστορικό του. Αυτή η δομική ιδιότητα επιτρέπει στο μοντέλο να κατανοεί εγγενώς τις μακρο-συνθετικές σχέσεις. Μπορεί να ανακαλέσει ένα μελωδικό μοτίβο που εισήχθη στα πρώτα δευτερόλεπτα ενός κομματιού και να αναπαράγει μαθηματικά την ακριβή αρμονική του ανάλυση ή το μελωδικό του θέμα (hook) δεκάδες δευτερόλεπτα αργότερα, διασφαλίζοντας τη συνθετική συνέχεια σε εκτεταμένους χρονικούς ορίζοντες.

- Απότομη Χρονική Κβαντοποίηση και Ευκρινείς Ενάρξεις Νότων

Επειδή το μοντέλο βελτιστοποιείται μέσω της κατηγορικής απώλειας διασταυρούμενης εντροπίας (categorical cross-entropy loss) σε σχέση με ένα διακριτό λεξιλόγιο του βιβλίου κωδικών (codebook), οι αποφάσεις του είναι εξαιρετικά εντοπισμένες και αποφασιστικές. Αυτή η δομική μορφή επιτρέπει στα αυτοπαλίνδρομα συστήματα να υπερέχουν στη σύνθεση κρουστικών μεταβατικών στοιχείων (percussive transients), ευκρινών χτυπημάτων από τύμπανα και καθαρών ενάρξεων (onsets) για τις νότες των οργάνων. Το μοντέλο δεν υποφέρει από χωρική ή χρονική θόλωση (blurring). Επιλέγοντας στο στοχαστικό του πλέγμα έναν οριστικό δείκτη, τοποθετεί τη νότα ακριβώς επάνω στο ρυθμικό πλέγμα (grid), οδηγώντας σε καθαρή, δυναμική ρυθμική κίνηση.

2) Μειονεκτήματα

- Το Σημείο Συμφόρησης του Inference (Αργή Γραμμική Παραγωγή)

Το καθοριστικό λειτουργικό ελάττωμα οποιουδήποτε αυτοπαλίνδρομου μοντέλου είναι ο αυστηρά διαδοχικός βρόχος συμπερασμού (sequential inference loop). Ενώ η εκπαίδευση μπορεί να παραλληλιστεί πλήρως σε όλη τη διάσταση του χρόνου με τη χρήση αιτιακών масκών, η παραγωγή κατά τον χρόνο εκτέλεσης (runtime) στερείται αυτής της δυνατότητας. Για την κατασκευή μιας ακολουθίας ήχου, το μοντέλο πρέπει να εκτελέσει ένα πλήρες πέρασμα προς τα εμπρός (forward pass) μέσα από δισεκατομμύρια παραμέτρους του Transformer για να προβλέψει ένα και μόνο χρονικό βήμα. Στη συνέχεια, προσαρτά αυτό το token στο παράθυρο του ιστορικού πλαισίου και επανεκτελεί ολόκληρο το δίκτυο για να προβλέψει το αμέσως επόμενο βήμα. Για παράδειγμα, για ένα τυπικό μουσικό κλιπ διάρκειας 30 δευτερολέπτων, το οποίο υποβάλλεται σε επεξεργασία με ρυθμό συμπίεσης πλαισίων RVQ της τάξης των 50Hz, ο υπολογιστικός κύκλος διαμορφώνεται ως:

$$30 \text{ sec} \times 50 \text{ frames/sec} = 1500 \text{ μεμονωμένα περάσματα forward}$$

- Αυτό δημιουργεί μια έντονη υπολογιστική συμφόρηση, καθώς ο χρόνος εκτέλεσης κλιμακώνεται γραμμικά —δηλαδή με πολυπλοκότητα $O(S)$ — σε σχέση με τη διάρκεια του κομματιού. Η διαδραστική σύνθεση σε πραγματικό χρόνο ή η ταχύτατη παράλληλη παραγωγή πολλαπλών μουσικών εναλλακτικών (variations) καθίσταται υπολογιστικά εξαιρετικά δυσχερές σε τυπικό υλικό hardware.
- Σύνθετα Σφάλματα και Μεροληψία Έκθεσης (Exposure Bias)
Κατά τη φάση της εκπαίδευσης, το μοντέλο λαμβάνει πάντα ως πλαίσιο εισόδου την πραγματική, έγκυρη ιστορία των δεδομένων (ground-truth history), μια στρατηγική γνωστή ως εξαναγκασμός από τον δάσκαλο (teacher forcing), βάσει του θεωρητικού μοντέλου των [Williams & Zipser \(1989\)](#). Ωστόσο, κατά τη διάρκεια του συμπερασμού (inference), το μοντέλο λειτουργεί σε συνθήκες απομόνωσης:

αναγκάζεται να τροφοδοτήσει τα δικά του προβλεπόμενα διακριτικά πίσω στο παράθυρο του περιβάλλοντος ως τη μοναδική «αλήθεια» για το επόμενο βήμα. Εάν το μοντέλο διαπράξει ένα μικρό στατιστικό λάθος στο χρονικό πλαίσιο t , όπως η επιλογή ενός διακριτικού που εισάγει μια ανεπαίσθητη απόκλιση στο τονικό ύψος (pitch drift), το σφάλμα αυτό κωδικοποιείται μόνιμα στο ιστορικό του. Καθώς η ακολουθία προχωρά, τα επόμενα βήματα υπολογίζουν τις εξαρτήσεις της προσοχής επάνω σε αυτή την εσφαλμένη βάση, οδηγώντας σε συσσωρευτικά σφάλματα (compounding errors). Σε εκτεταμένα χρονοδιαγράμματα παραγωγής, αυτή η μεροληψία έκθεσης (exposure bias) προκαλεί συχνά τη δομική μετατόπιση (structural drift) του συστήματος. Η μουσική έξοδος ενδέχεται σταδιακά είτε να διαλυθεί σε χαοτικό ψηφιακό θόρυβο είτε να απομακρυνθεί εντελώς από τις σημασιολογικές οδηγίες της αρχικής κειμενικής προτροπής.

4.2 Μοντέλα Λανθάνουσας Διάχυσης: Συνεχής Ακουστική Αποθορυβοποίησης σε Λανθάνοντες Χώρους

Σε άμεση αντίθεση με το παράδειγμα της διακριτής, συμβολικής πρόβλεψης ακολουθιών που χαρακτηρίζει την Αυτοπαλίνδρομη Μοντελοποίηση Γλώσσας, το μοντέλο Λανθάνουσας Διάχυσης (Latent Diffusion Model - LDM) αντιμετωπίζει τη σύνθεση κειμένου σε ήχο ως μια συνεχή, χωροχρονική εργασία βελτιστοποίησης, όπως τεκμηριώνεται στα έργα των [Rombach et al. \(2022\)](#) και [Liu et al. \(2023\)](#). Η προσέγγιση αυτή, η οποία εισήχθη πρωτοποριακά στον οπτικό τομέα και προσαρμόστηκε με επιτυχία σε ακουστικά δεδομένα από αρχιτεκτονικά πλαίσια όπως το AudioLDM ([Liu et al., 2023](#)) και το Stable Audio ([Evans et al., 2024](#)), παρακάμπτει πλήρως τα διακριτικά στοιχεία (tokens) των βιβλίων κωδικών.

Αντίθετα, τα LDM λειτουργούν μέσα σε έναν συνεχή, χαμηλής διαστατικότητας λανθάνοντα χώρο (continuous low-dimensional latent space). Αυτός ο χώρος υλοποιείται συνήθως μέσω ενός βελτιστοποιημένου καναλιού Μεταβλητού Αυτοκωδικοποιητή (Variational Autoencoder - VAE) ή ενός πίνακα συντεταγμένων Log-Mel Φασματογραφήματος. Τα

Διπλωματική Εργασία

μοντέλα αυτά αναπαριστούν την κατανομή-στόχο μαθαίνοντας να αντιστρέφουν μια στοχαστική διαδικασία αλλοίωσης πολλαπλών βημάτων.

4.2.1 Μαθηματικό Πλαίσιο και η Τροχιά Αποθορυβοποίησης

Η εκτέλεση μιας αρχιτεκτονικής λανθάνουσας διάχυσης χωρίζεται σε δύο ασύμμετρες φάσεις: μια εμπρός διαδρομή αλλοίωσης Μαρκοβιανού τύπου (forward corruption process) και μια βαθιά νευρωνική αντίστροφη διαδρομή αποθορυβοποίησης (reverse denoising process), σύμφωνα με τη θεωρητική θεμελίωση των [Ho et al. \(2020\)](#).

Έστω $\mathbf{z}_0 \in \mathbb{R}^{F \times T}$ η απαλλαγμένη από αλλοιώσεις, συνεχής λανθάνουσα αναπαράσταση ενός ηχητικού σήματος, όπου η μεταβλητή F ορίζει τον αριθμό των συχνοτικών ζωνών (frequency bins) και η μεταβλητή T αντιπροσωπεύει τη συμπιεσμένη χρονική διάσταση του πλαισίου.

1) Η Εμπρός Διαδικασία (Διαφθορά Δεδομένων - Forward Process)

Η εμπρός διαδικασία υποβαθμίζει προοδευτικά τη δομική ακεραιότητα της ακουστικής αναπαράστασης, εισάγοντας μικροσκοπικά βήματα λευκού θορύβου Gauss κατά μήκος μιας διακριτής χρονικής κλίμακας $t \in [1, T]$. Η διαδικασία αυτή διέπεται από ένα προκαθορισμένο χρονοδιάγραμμα διακύμανσης (variance schedule) $\beta_1, \beta_2, \dots, \beta_T$ όπως φαίνεται στον παρακάτω τύπο:

$$q(\mathbf{z}_t \vee \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1 - \beta_t} \mathbf{z}_{t-1}, \beta_t \mathbf{I}) \quad (61)$$

Μια κρίσιμη μαθηματική ιδιότητα αυτής της διατύπωσης είναι ότι η κατάσταση σε οποιοδήποτε αυθαίρετο χρονικό βήμα t μπορεί να δειγματοληπτηθεί απευθείας από τον αρχικό, καθαρό τανυστή $\mathbf{z}_0$, χωρίς να απαιτείται η διαδοχική εκτέλεση όλων των ενδιάμεσων βημάτων:

$$q(\mathbf{z}_t \vee \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t} \mathbf{z}_0, (1 - \bar{\alpha}_t) \mathbf{I}) \quad (62)$$

Όπου

$$\alpha_t = 1 - \beta_t \quad \text{και} \quad \bar{\alpha}_t = \prod_{i=1}^t \alpha_i \quad (63)$$

Καθώς η χρονική μεταβλητή πλησιάζει στο όριο $t \rightarrow T$, τα αρχικά ηχητικά χαρακτηριστικά διαγράφονται ολοκληρωτικά, μετατρέποντας τον τανυστή z_T σε καθαρό, χαοτικό λευκό θόρυβο Gauss.

2) Η Αντίστροφη Διαδικασία (Υπό Συνθήκη Σύνθεση - Reverse Process)

Ο στόχος του νευρωνικού δικτύου είναι να εκμάθει με ακρίβεια την αντίστροφη τροχιά: μια παραμετροποιημένη κατανομή πιθανότητας $p_\theta(z_{t-1}, z_t)$, η οποία είναι ικανή να αφαιρεί επαναληπτικά συγκεκριμένα κλάσματα θορύβου, αποκαλύπτοντας σταδιακά τα υποκείμενα ακουστικά χαρακτηριστικά. Αυτή η αντίστροφη πορεία καθοδηγείται και περιορίζεται από τις πυκνές σημασιολογικές ενσωματώσεις κειμένου H_{text} που εξάγονται από το γλωσσικό front-end:

$$p_\theta(z_{t-1} | z_t, H_{text}) = N(z_{t-1}; \mu_\theta(z_t, t, H_{text}), \Sigma_\theta(z_t, t, H_{text})) \quad (64)$$

Για την εκτέλεση αυτού του σύνθετου υπολογισμού επιστρατεύεται μια βαθιά νευρωνική ραχοκοκαλιά —παραδοσιακά ένα συνελκτικό δίκτυο U-Net ([Ronneberger et al., 2015](#)) ή ένας Μετασχηματιστής Διάχυσης (Diffusion Transformer - DiT) ([Peebles & Xie, 2023](#))— η οποία έχει βελτιστοποιηθεί με σκοπό να προβλέπει το ακριβές προφίλ του θορύβου ϵ που εμπεριέχεται στον αλλοιωμένο τανυστή κατά τη χρονική στιγμή t . Το δίκτυο ενσωματώνει τα διανυσματικά χαρακτηριστικά του κειμένου απευθείας στα εσωτερικά του επίπεδα επεξεργασίας μέσω μηχανισμών διασταυρούμενης προσοχής (cross-attention), ακολουθώντας τη μαθηματική δομή του [Vaswani \(2017\)](#). Με αυτόν τον τρόπο,

αναδιαμορφώνει δυναμικά το πεδίο εκτίμησης του θορύβου με βάση τις γλωσσικές οδηγίες της ανθρώπινης προτροπής, διασφαλίζοντας υψηλή ακουστική συμμόρφωση.

4.2.2 Αναλυτική Αξιολόγηση: Πλεονεκτήματα και Μειονεκτήματα

Η μετάβαση από τα αυτοπαλίνδρομα συστήματα στη μοντελοποίηση λανθάνουσας διάχυσης (Latent Diffusion Models - LDMs) αναδιαμόρφωσε πλήρως το τοπίο της διατροφικής παραγωγής ήχου. Τα συστήματα αυτά καθιερώθηκαν ως μια εξαιρετικά κλιμακώσιμη (scalable) εναλλακτική λύση, χάρη σε μια σειρά από διακριτά αρχιτεκτονικά και υπολογιστικά πλεονεκτήματα. Παράλληλα, όμως, εισήγαγαν νέες προκλήσεις όσον αφορά την αυστηρή μουσική δομή, όπως επισημαίνεται στα έργα των [Liu et al. \(2023\)](#) και [Evans et al. \(2024\)](#).

1) Πλεονεκτήματα

- Μαζικά Επιταχυνόμενη, Παράλληλη Δημιουργία (Υπολογιστική Πολυπλοκότητα $O(1)$)

Το σημαντικότερο μηχανικό πλεονέκτημα του παραδείγματος LDM έναντι των αυτοπαλίνδρομων δομών είναι η πλήρης ανεξαρτησία του από τα διαδοχικά (step-by-step) χρονικά σημεία συμφόρησης. Επειδή τα μοντέλα διάχυσης αντιμετωπίζουν τον λανθάνοντα χώρο του ήχου ως έναν ενοποιημένο, πολυδιάστατο καμβά, έχουν τη δυνατότητα να υπολογίζουν τις ενημερώσεις σε όλα τα χρονικά ηχητικά πλαίσια ταυτόχρονα (παράλληλα) κατά τη διάρκεια μιας διέλευσης προς τα εμπρός (forward pass), σύμφωνα με το θεωρητικό πλαίσιο των [Rombach et al. \(2022\)](#). Όταν επιστρατεύονται επιταχυνόμενοι αριθμητικοί χρονοπρογραμματιστές, όπως το πλαίσιο DDIM ([Song et al., 2022](#)) ή το DPM-Solver++ ([Lu et al., 2022](#)), ο βρόχος του συμπερασμού (inference loop) απαιτεί έναν αυστηρά σταθερό αριθμό βημάτων (π.χ. 15 έως 50 επαναλήψεις), ο οποίος είναι πλήρως ανεξάρτητος από την πραγματική χρονική διάρκεια του κομματιού που συντίθεται.

Η παραγωγή ενός ακουστικού αποσπάσματος διάρκειας 10 δευτερολέπτων έναντι ενός κομματιού 60 δευτερολέπτων απαιτεί ακριβώς τον ίδιο αριθμό εκτελέσεων του δικτύου. Αυτή η δυνατότητα καθολικής παράλληλης επεξεργασίας επιτρέπει στα μοντέλα διάχυσης να επιτυγχάνουν ταχύτητες παραγωγής σε υπο-πραγματικό χρόνο (*sub-real-time*), ξεκλειδώνοντας τη δυνατότητα για διαδραστικές εφαρμογές σύνθεσης μεγάλης κλίμακας.

- Απουσία Μεροληψίας Έκθεσης (Exposure Bias) και Σφαλμάτων Σύνθεσης

Επειδή τα LDM δεν βασίζονται στην επαναληπτική τροφοδοσία των δικών τους προηγούμενων προγνώσεων σε ένα συνεχώς επεκτεινόμενο παράθυρο ιστορικού, είναι εντελώς άτρωτα στο φαινόμενο της μεροληψίας έκθεσης (*exposure bias*) που υποβαθμίζει τις αυτοπαλίνδρομες αλυσίδες. Όπως τεκμηριώνεται από τους [Ho et al. \(2020\)](#), εάν το δίκτυο αποθορυβοποίησης υπολογίσει μια ατελή ή ελαφρώς εσφαλμένη πρόβλεψη θορύβου στο χρονικό βήμα t , το συνολικό σφάλμα δεν συσσωρεύεται εκθετικά. Αντίθετα, αναδιανέμεται φυσικά και αυτοδιορθώνεται κατά τη διάρκεια των εναπομεινάντων βημάτων της αντίστροφης τροχιάς. Ως εκ τούτου, το μοντέλο δεν παρουσιάζει ποτέ ξαφνική δομική κατάρρευση, ούτε μετατοπίζεται σε στατικό ψηφιακό θόρυβο, διασφαλίζοντας κορυφαία ακουστική σταθερότητα.

- Υψηλή Αντιληπτική Πιστότητα και Οργανικές Ακουστικές Υφές

Λειτουργώντας μέσα σε έναν συνεχή λανθάνοντα χώρο κινητής υποδιαστολής (*continuous floating-point latent space*) αντί για ένα κβαντισμένο ακέραιο πλέγμα δεικτών, τα LDM υπερέχουν στη μοντελοποίηση ρευστών, πυκνών και σύνθετων φασμάτων συχνότητας ([Liu et al., 2023](#)). Αυτή η συνεχής τοπολογία αποτρέπει την εισαγωγή διακριτών τεχνουργημάτων στις ακμές των αναπαραστάσεων. Το γεγονός αυτό καθιστά τις αρχιτεκτονικές διάχυσης εξαιρετικά ικανές στη σύνθεση πλούσιων

και εξελισσόμενων ατμοσφαιρικών περιβαλλόντων (ambient pads), πυκνών ορχηστρικών εγχόρδων (strings) και φυσικών φωνητικών ηχοχρωμάτων και πολύπλοκων στερεοφωνικών περιβαλλόντων αντήχησης (reverberation) που απαιτούν ομαλές, συνεχείς ακουστικές καμπύλες.

2) Μειονεκτήματα

- Ασθενής Μακροδομικός Έλεγχος και Ρυθμική Μετατόμιση (Rhythm Drift)

Σε αντίθεση με τα αυτοπαλίνδρομα μοντέλα, τα οποία επιστρατεύουν μια αυστηρή αιτιακή μάσκα (causal mask) για να κλειδώσουν τις χρονικές εξαρτήσεις προς τα πίσω, ένα τυπικό μοντέλο διάχυσης βασίζεται σε ένα καθολικό, αμφίδρομο προφίλ προσοχής (global bidirectional attention profile). Ενώ η προσέγγιση αυτή είναι ιδανική για την καταγραφή ολιστικών υφών και ηχοχρωμάτων, δυσκολεύεται να επιβάλει ένα άκαμπτο, μαθηματικά ακριβές ρυθμικό πλέγμα (click-track) για εκτεταμένες χρονικές περιόδους. Επειδή το νευρωνικό δίκτυο βελτιστοποιεί την παγκόσμια δομική ευθυγράμμιση επάνω στον καμβά αντί για μια αυστηρή χρονογραμμή βήμα προς βήμα, ο ήχος που παράγεται από μοντέλα διάχυσης υποφέρει συχνά από ρυθμική μετατόπιση / διολίσθηση (rhythm drift). Για παράδειγμα ένα σταθερό μοτίβο τυμπάνων (drum loop) που ξεκινά τέλεια στην αρχή ενός κομματιού ενδέχεται να παρουσιάσει ανεπαίσθητες χρονικές μετατοπίσεις μικροδευτερολέπτων (micro-timing deviations) καθώς προχωρά η παραγωγή. Επίσης, αυτές οι αποκλίσεις συσσωρεύονται, προκαλώντας ασύμμετρη ευθυγράμμιση με τον μετρονόμο, ενώ παράλληλα «θολώνουν» και διαλύουν τη γενικότερη μακροδομή του κομματιού, όπως τις σαφείς δομικές αλλαγές των συγχορδιών (chord progressions) ή τις καθορισμένες μεταβάσεις μεταξύ των μουσικών μερών.

- Εξασθενημένα Μεταβατικά Φαινόμενα (Softened Transients) και θολές ενάρξεις νοτών

Επειδή η αντικειμενική συνάρτηση απώλειας των μοντέλων διάχυσης βελτιστοποιεί το Μέσο Τετράγωνο Σφάλμα (Mean Squared Error - MSE) επάνω σε συνεχείς μορφές και γεωμετρίες του φασματογραφήματος, το δίκτυο εμφανίζει μια εγγενή στατιστική προκατάληψη προς τον υπολογισμό μιας μέσης κατανομής των τιμών ([Ho et al., 2020](#)). Ακουστικά, η μαθηματική αυτή ιδιότητα εκδηλώνεται ως μια αισθητή απώλεια ακαριαίας ενέργειας κατά τη διάρκεια αιχμηρών, γρήγορων μεταβατικών φαινομένων (transients). Τα κρουστικά χτυπήματα —όπως η καθαρή και κοφτή επίθεση (attack) ενός ταμπούρου ή το ακαριαίο τράβηγμα (plucking) των χορδών μιας ακουστικής κιθάρας— συχνά χάνουν την άμεση, δυναμική τους επίδραση (punch) σε έναν αγωγό διάχυσης. Η έναρξη της νότας (note onset) εμφανίζεται «μουτζουρωμένη» ή «μαλακωμένη» (smeared), καθώς οι τιμές διαχέονται στα γειτονικά χρονικά πλαίσια. Το φαινόμενο αυτό καθιστά υποχρεωτική τη χρήση βαριάς επεξεργασίας στο back-end, μέσω εξειδικευμένων αντιπαραθετικών νευρωνικών vocoders όπως η αρχιτεκτονική HiFi-GAN ([Kong et al., 2020](#)), προκειμένου να αποκατασταθεί πλήρως η δυναμική, η καθαρότητα και ο ορισμός των κρουστικών σημάτων στο τελικό αρχείο κυματομορφής.

5. Εμπειρικά Σημεία Αναφοράς & Μεθοδολογία Αξιολόγησης

5.1 Αντικειμενικές Μετρήσεις Αξιολόγησης

Τα διατροφικά συστήματα παραγωγής ήχου προϋποθέτουν μετρήσεις αξιολόγησης που είναι πλήρως ανεξάρτητες από την ανθρώπινη υποκειμενικότητα, αλλά και απαλλαγμένες από τους περιορισμούς των παραδοσιακών συναρτήσεων απώλειας. Όπως επισημάνθηκε, η χρήση μετρικών όπως το Μέσο Τετραγωνικό Σφάλμα (MSE) απευθείας επάνω στα pixel του φασματογραφήματος αποτυγχάνει, καθώς τιμωρεί αυστηρά τις ανεπαίσθητες μετατοπίσεις φάσης ή χρόνου, ακόμη και όταν δύο ηχητικά σήματα ακούγονται απολύτως πανομοιότυπα. Για την υπέρβαση αυτού του προβλήματος, τα σύγχρονα κριτήρια αξιολόγησης (benchmarks) βασίζονται σε δύο προηγμένες, χωρίς αναφορά (reference-free) αντικειμενικές μετρήσεις. Οι μετρήσεις αυτές αναλύουν την ποιότητα της παραγωγής μέσα από παγωμένους (frozen), προεκπαιδευμένους χώρους ενσωματώσεων (embedding spaces) και είναι η Απόσταση Ήχου Fréchet (Fréchet Audio Distance - FAD) για την αξιολόγηση της ακουστικής πιστότητας και η Βαθμολογία CLAP (CLAP Score) για τη μέτρηση της σημασιολογικής ευθυγράμμισης μεταξύ κειμένου και ήχου.

5.1.1 Απόσταση Ήχου Fréchet (Fréchet Audio Distance - FAD)

Η Απόσταση Ήχου Fréchet (FAD) αποτελεί την προσαρμογή της επιτυχημένης μετρικής FID (Fréchet Inception Distance) των εικόνων στο ακουστικό πεδίο, όπως εισηγήθηκαν οι [Kilgour et al. \(2019\)](#). Αντί να συγκρίνει μεμονωμένα δείγματα (sample-to-sample), η FAD αξιολογεί τη στατιστική απόσταση ανάμεσα σε ολόκληρες κατανομές αρχείων ήχου, καθιστώντας την εξαιρετικά ανθεκτική σε μικρομετατοπίσεις και ευθυγραμμίσεις φάσης.

Η διαδικασία υλοποιείται μέσω ενός παγωμένου, προεκπαιδευμένου δικτύου εξαγωγής ακουστικών χαρακτηριστικών, όπως το VGGish ή τα PANNs (Pretrained Audio Neural Networks) που αναπτύχθηκαν από τους [Kong et al. \(2020\)](#):

1. Το δίκτυο δέχεται ένα μεγάλο σύνολο αναφοράς από πραγματικά αρχεία ήχου (background/real data) και εξάγει τα αντίστοιχα διανύσματα ενσωμάτωσης (embeddings). Η κατανομή τους μοντελοποιείται ως μια πολυμεταβλητή Γκαουσιανή (multivariate Gaussian): $N(\mu_r, \Sigma_r)$.
2. Στη συνέχεια, η ίδια διαδικασία επαναλαμβάνεται για το σύνολο των παραχθέντων (συνθετικών) αρχείων ήχου από το προς αξιολόγηση μοντέλο, διαμορφώνοντας την κατανομή: $N(\mu_g, \Sigma_g)$.

Στο επίπεδο της μαθηματικής τυποποίησης, η FAD ορίζεται ως η απόσταση Wasserstein-2 μεταξύ των δύο αυτών Γκαουσιανών κατανομών:

$$d^2(N_r, N_g) = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (65)$$

Όπου:

- $\|\mu_r - \mu_g\|_2^2$: Η τετραγωνισμένη Ευκλείδεια απόσταση μεταξύ των μέσων διανυσμάτων των δύο κατανομών.
- Σ_r, Σ_g : Οι πίνακες συνδιακύμανσης για το πραγματικό και το συνθετικό σύνολο δεδομένων αντίστοιχα.
- Tr : Ο τελεστής ίχνους πίνακα (trace), ο οποίος επιστρέφει το άθροισμα των διαγώνιων στοιχείων της προκύπτουσας μήτρας.

Η ερμηνεία των αποτελεσμάτων της απόστασης Fréchet, ακολουθεί ως εξής:

- FAD $\rightarrow 0$ (Χαμηλή Βαθμολογία): Υποδηλώνει ότι η στατιστική κατανομή των παραχθέντων δειγμάτων προσεγγίζει στενά την κατανομή των πραγματικών ήχων. Αυτό μεταφράζεται σε υψηλή ακουστική πιστότητα, φυσικό ηχόχρωμα και απουσία

ψηφιακών τεχνουργημάτων ή παραμορφώσεων, αποτυπώνοντας την αντιληπτική ποιότητα με τρόπο που συνάδει με την ανθρώπινη ακοή.

- Υψηλή Βαθμολογία FAD: Αποτελεί σαφή ένδειξη σοβαρής ακουστικής υποβάθμισης. Ακόμη και στην περίπτωση που η μουσική σύνθεση είναι τεχνικά ή θεωρητικά ακριβής, μια υψηλή τιμή FAD αποδεικνύει την ύπαρξη αφύσικων ακουστικών στοιχείων. Τέτοια προβλήματα περιλαμβάνουν τον ψηφιακό βόμβο (digital buzz), τις ακυρώσεις φάσης (phase cancellations), τα έντονα τεχνουργήματα (artifacts) ή τις πεπλατυσμένες συχνοτικές ζώνες που εισάγονται από τον παραγωγικό αγώνα ή τα επίπεδα του νευρωνικού αποκωδικοποιητή.

Η FAD υπερέχει των παραδοσιακών μετρικών επειδή ευθυγραμμίζεται στενά με την ανθρώπινη ακουστική αντίληψη. Καταγράφει την ολιστική ποιότητα και την καθαρότητα μιας παραγωγής, αγνοώντας τις ασήμαντες για το ανθρώπινο αυτί μικρομετατοπίσεις στο χρονικό πεδίο, οι οποίες υπό άλλες συνθήκες θα αλίωναν πλήρως τα αποτελέσματα των τυπικών υπολογισμών σφάλματος.

5.1.2 Βαθμολογία CLAP (Contrastive Language-Audio Pretraining Score)

Ενώ η FAD μετρά την καθαρή ακουστική ποιότητα, δεν παρέχει καμία πληροφορία σχετικά με το αν ο παραγόμενος ήχος ανταποκρίνεται στις οδηγίες του χρήστη. Για τη μέτρηση της διατροφικής, σημασιολογικής συνέπειας επιστρατεύεται η Βαθμολογία CLAP (CLAP Score), η οποία βασίζεται σε μοντέλα Αντιπαραθετικής Προεκπαίδευσης Γλώσσας-Ήχου, όπως τεκμηριώνεται στα έργα των [Elizalde et al. \(2023\)](#) και [Wu et al. \(2023\)](#).

Η αρχιτεκτονική CLAP αποτελείται από δύο ανεξάρτητα υποδίκτυα: έναν κωδικοποιητή κειμένου (Text Encoder) και έναν κωδικοποιητή ήχου (Audio Encoder). Τα υποδίκτυα αυτά έχουν εκπαιδευτεί αντιπαραθετικά (contrastively) σε τεράστια σύνολα δεδομένων, ώστε να προβάλλουν τις κειμενικές περιγραφές και τα αντίστοιχα ακουστικά σήματα σε έναν κοινό λανθάνοντα χώρο ενσωματώσεων (shared embedding space).

Κατά την αξιολόγηση, η γλωσσική προτροπή (prompt) του χρήστη διοχετεύεται στον κωδικοποιητή κειμένου, παράγοντας το διάνυσμα $e_t = E_{text}(\text{prompt})$. Ταυτόχρονα, ο συντεθημένος ήχος που παρήγαγε η γεννητική μηχανή διοχετεύεται στον κωδικοποιητή ήχου, παράγοντας το διάνυσμα $e_a = E_{audio}(\text{audio})$. Και τα δύο διανύσματα κανονικοποιούνται ώστε να έχουν μοναδιαίο μήκος ($\|e\|_2 = 1$).

Στη σύγχρονη βιβλιογραφία, η Βαθμολογία CLAP ορίζεται ως η ομοιότητα συνημιτόνου (cosine similarity) ανάμεσα στις δύο αυτές διατροπικές αναπαραστάσεις:

$$CLAP\ Score = \cos(e_t, e_a) = \frac{e_t \cdot e_a}{\|e_t\|_2 \|e_a\|_2} = e_t^T e_a \quad (66)$$

Η ερμηνεία των αποτελεσμάτων της βαθμολογίας CLAP, ακολουθεί ως εξής:

- CLAP $\rightarrow 1$ (Υψηλή Τιμή): Αντιπροσωπεύει μια ισχυρή σημασιολογική συσχέτιση μεταξύ της κειμενικής περιγραφής και του παραχθέντος ακουστικού σήματος. Τα φυσικά ακουστικά χαρακτηριστικά που εντοπίζονται στο παραχθέν ηχητικό κομμάτι (όπως το τέμπο, η ενορχήστρωση, τα ενεργά όργανα και οι περιβαλλοντικές ιδιότητες) εναρμονίζονται απόλυτα με τις γλωσσικές συνθήκες που τέθηκαν στη συμβολοσειρά του κειμένου.
- CLAP $\rightarrow 0$ ή Αρνητικές Τιμές: Αποδεικνύει την πλήρη κατάρρευση της συσχέτισης μεταξύ της κειμενικής περιγραφής και του ακουστικού σήματος. Η γεννητική μηχανή αγνόησε τους περιορισμούς της κειμενικής προτροπής, συνθέτοντας έναν αυθαίρετο ήχο που στερείται οποιασδήποτε στατιστικής ή σημασιολογικής συσχέτισης με την προτροπή του χρήστη.

Η βαθμολογία CLAP αποτελεί το απαραίτητο συμπλήρωμα της FAD. Ενώ η FAD διασφαλίζει ότι η έξοδος ακούγεται ως πραγματική, υψηλής πιστότητας μουσική, η CLAP εγγυάται ότι η μουσική αυτή ανταποκρίνεται με ακρίβεια στις εντολές και το εννοιολογικό πλαίσιο που καθορίζει ο άνθρωπος.

5.2 Υποκειμενικές Μετρήσεις Αξιολόγησης: Μέση Βαθμολογία Γνώμης (MOS)

Ενώ οι αντικειμενικές στατιστικές μετρήσεις, όπως η Απόσταση Ήχου Fréchet (Fréchet Audio Distance - FAD) και η Βαθμολογία CLAP, παρέχουν αναπαραγώγιμα σημεία αναφοράς για την αξιολόγηση αλγορίθμων μηχανικής μάθησης, αδυνατούν να αντικαταστήσουν πλήρως την ανθρώπινη αντίληψη, όπως υπογραμμίζουν οι [Kilgour et al. \(2019\)](#) και [Elizalde et al. \(2023\)](#). Τα αντικειμενικά μοντέλα επεξεργάζονται τα δεδομένα μέσω σταθερών νευρωνικών δικτύων, τα οποία μπορούν να παρακαμφθούν συστηματικά ή να «ξεγελαστούν».

Για παράδειγμα, ένα παραγωγικό μοντέλο μπορεί να δημιουργήσει δομικά ασύνδετη ή αισθητικά δυσάρεστη μουσική, η οποία όμως επιτυγχάνει υψηλή βαθμολογία CLAP απλώς και μόνο επειδή εμφανίζει υπερβολική ευρετηρίαση (over-indexing) σε συγκεκριμένα ακουστικά στοιχεία-κλειδιά που σχετίζονται με την προτροπή. Για την αξιολόγηση της πραγματικής καλλιτεχνικής αξίας, της ροής των μεταβάσεων και του συναισθηματικού συντονισμού της συνθετικής μουσικής, η διατροπική έρευνα βασίζεται σε Υποκειμενικές Μετρήσεις Αξιολόγησης (Subjective Evaluation Metrics), με κεντρικό πυλώνα το παράδειγμα της Μέσης Βαθμολογίας Γνώμης (Mean Opinion Score - MOS), που θεσπίστηκε από την [ITU-T \(1996\)](#).

5.2.1 Μηχανική και Διαδικασία των Δοκιμών MOS

Η Μέση Βαθμολογία Γνώμης είναι μια ψυχοφυσική μετρική που υιοθετήθηκε αρχικά από τη μηχανική τηλεπικοινωνιών και τη σύνθεση ομιλίας ([ITU-T, 1996](#)). Σε μια διαδικασία αξιολόγησης συστημάτων text-to-audio, σχεδιάζεται και αναπτύσσεται μια τυποποιημένη δοκιμασία ακρόασης (listening test), η οποία υποβάλλεται σε μια επιλεγμένη ομάδα ανθρώπινων αξιολογητών. Η ομάδα αυτή αποτελείται συνήθως από δύο ξεχωριστές κατηγορίες:

1. Ανεξάρτητους αξιολογητές (crowdsourced raters): Προέρχονται από το ευρύ κοινό (π.χ. μέσω πλατφορμών όπως το Amazon Mechanical Turk).
2. Εμπειρογνώμονες (experts): Μουσικολόγους ή μηχανικούς ήχου που είναι σε θέση να ανιχνεύσουν ανεπαίσθητες ακουστικές παραμορφώσεις και αποκλίσεις φάσης.

Κατά τη διάρκεια μιας επίσημης δοκιμασίας ακρόασης, στους αξιολογητές παρουσιάζεται μια τυχαιοποιημένη, διπλά τυφλή (double-blind) σειρά παραγόμενων ηχητικών δειγμάτων μαζί με τις αντίστοιχες κειμενικές προτροπές. Οι αξιολογητές ακούν κάθε κλιπ υπό ελεγχόμενες ακουστικές συνθήκες και του αποδίδουν μια διακριτή αριθμητική βαθμολογία με βάση μια κλίμακα Likert 5 βαθμών, όπως φαίνεται στο παρακάτω :

Βαθμολογία	Ποιοτική Κατηγορία	Αντιληπτική Κατάσταση
5	Άριστα(Excellent)	Καμία αισθητή υποβάθμιση ή απόκλιση
4	Καλό(Good)	Αισθητή αλλά μη ενοχλητική υποβάθμιση
3	Μέτριο(Fair)	Μέτρια ενοχλητική υποβάθμιση ή μικρές σημασιολογικές παραλείψεις
2	Κακό(Bad)	Σοβαρά ενοχλητικά τεχνουργήματα ή σημαντική δομική ασυνέπεια
1	Πολύ κακό(Fail)	Πλήρης ακουστική κατάρρευση ή εντελώς άσχετη έξοδος

Πίνακας 1: Πίνακας Βαθμολογίας Likert-Αξιολόγηση MOS

Μαθηματικά, η συνολική Μέση Βαθμολογία Γνώμης (MOS) για μια συγκεκριμένη διαμόρφωση γενετικού μοντέλου υπολογίζεται ως ο αριθμητικός μέσος όρος του συνόλου των μεμονωμένων αξιολογήσεων:

$$MOS = \frac{1}{N} \sum_{i=1}^N R_i \quad (67)$$

Όπου η μεταβλητή $R_i \in \{1,2,3,4,5\}$ αντιπροσωπεύει τη διακριτή βαθμολογία που αποδόθηκε από τον αξιολογητή i , και η μεταβλητή N ορίζει τον συνολικό αθροιστικό αριθμό μεμονωμένων αξιολογήσεων που καταγράφηκαν για το συγκεκριμένο δείγμα.

5.2.2 Υποκειμενική Αξιολόγηση Διπλού Άξονα

Για την ουσιαστική αξιολόγηση των συστημάτων μετατροπής κειμένου σε μουσική, μια ενιαία, μονοδιάστατη (flat) βαθμολογία MOS δεν κρίνεται επαρκής. Ένα μουσικό κομμάτι ενδέχεται να διαθέτει κορυφαία ποιότητα παραγωγής και ηχογράφησης, αλλά να αγνοεί πλήρως τις οδηγίες της προτροπής.

Προκειμένου να αποτυπωθεί η πολυδιάστατη φύση της παραγωγής, οι σύγχρονες μελέτες χωρίζουν τις υποκειμενικές δοκιμασίες σε δύο διακριτούς λειτουργικούς άξονες, σύμφωνα με το μεθοδολογικό πλαίσιο των [Copet et al. \(2023\)](#):

1) MOS-Q / MOS-M (Ακουστική Ποιότητα και Μουσικότητα)

Ο άξονας αυτός αξιολογεί τη φυσική ποιότητα του σήματος χαμηλού επιπέδου και τη γενικότερη μακροδομική σύνθεση της παραγόμενης κυματομορφής, ανεξάρτητα από το περιεχόμενο του κειμένου. Οι αξιολογητές καλούνται να απομονώσουν τις εξής ιδιότητες:

- Ηχοχρωματική Πιστότητα (Timbral Fidelity): Εξέταση του αν τα παραγόμενα όργανα ακούγονται φυσικά και αυθεντικά ή αν εμφανίζουν τεχνητές, μεταλλικές και ρομποτικές παραμορφώσεις.
- Δομική Σύνθεση (Structural Composition): Αξιολόγηση του αν η μουσική εξέλιξη ακολουθεί τις ανθρώπινες συνθετικές συμβάσεις (π.χ. λογική εναλλαγή συγχορδιών, διατήρηση σταθερού ρυθμού, φυσικές μεταβάσεις νοτών) ή αν αναλώνεται σε χαοτικές συστάδες συχνοτήτων.

- Ισορροπία Μίξης και Mastering (Mixing & Mastering Balance): Επαλήθευση του αν τα επίπεδα των οργάνων είναι καθαρά διαχωρισμένα στον στερεοφωνικό χώρο ή αν πάσχουν από καταστροφικές ακυρώσεις φάσης.

2) MOS-A / MOS-R (Συνάφεια και Στοιχίση Κειμένου)

Ο άξονας αυτός μετρά την ακρίβεια της σημασιολογικής ευθυγράμμισης που επιτυγχάνεται από τα επίπεδα διασταυρούμενης προσοχής (cross-attention). Οι βαθμολογητές μελετούν τη γλωσσική εντολή εισόδου και αξιολογούν σε ποιο βαθμό ο ήχος ανταποκρίνεται σε αυτήν, εξετάζοντας δύο επίπεδα λεπτομέρειας:

- Σαφής Αντιστοιχίση Περιεχομένου (Explicit Content Matching): Επαλήθευση του αν τα συγκεκριμένα ζητούμενα όργανα, το μουσικό είδος και το τέμπο είναι ακουστικά παρόντα (για παράδειγμα, αν ένα «σόλο ηλεκτρικής κιθάρας» συντέθηκε με ακρίβεια).
- Αντιστοιχίση Αφηρημένων Χαρακτηριστικών (Abstract Feature Matching): Αξιολόγηση του αν το μοντέλο αποτύπωσε με επιτυχία υφολογικούς προσδιορισμούς υψηλού επιπέδου, ιστορικές εποχές ή συναισθηματικά ηχοτοπία (για παράδειγμα, τη μετάφραση όρων όπως «ζοφερή ατμοσφαιρική lo-fi» ή «θριαμβευτικό ορχηστρικό κρεσέντο» στις σωστές ακουστικές τους αναπαραστάσεις).

Η αξιολόγηση διπλού άξονα MOS επιτρέπει στους ερευνητές να εντοπίσουν με ακρίβεια τις αδυναμίες ενός μοντέλου. Αν ένα σύστημα επιτυγχάνει υψηλό MOS-Q αλλά χαμηλό MOS-A, το πρόβλημα εντοπίζεται στον μηχανισμό της διατροπικής ευθυγράμμισης (cross-attention front-end). Αντίθετα, αν εμφανίζει υψηλό MOS-A αλλά χαμηλό MOS-Q, η αδυναμία έγκειται στον γενετικό πυρήνα αποθορυβοποίησης ή στον νευρωνικό αποκωδικοποιητή του back-end.

5.3 Υποκειμενικές Μετρήσεις Πολλαπλών Συνθηκών MUSHRA

Για την εκτέλεση εξαιρετικά ακριβών ακουστικών συγκρίσεων, οι αγωγοί αξιολόγησης επιστρατεύουν συχνά το αυστηρότερο πλαίσιο MUSHRA (Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems), το οποίο έχει θεσπιστεί βάσει του προτύπου ITU-R Recommendation [BS.1534 \(2001\)](#).

Σε αντίθεση με τη δοκιμή MOS, η οποία αξιολογεί τα δείγματα μεμονωμένα και ανεξάρτητα, μια δοκιμή MUSHRA βασίζεται σε ένα σχήμα άμεσης συγκριτικής αξιολόγησης.

Παρουσιάζει στον ανθρώπινο βαθμολογητή ένα σαφώς αναγνωρίσιμο, υψηλής πιστότητας και μη συμπιεσμένο κομμάτι αναφοράς (explicit reference), μαζί με μια σειρά από τυχαιοποιημένα δείγματα προς αξιολόγηση, όπως αναδεικνύεται στα έργα των [Zeghidour et al. \(2021\)](#) και [Copet et al. \(2023\)](#).

Μέσα στη στοίβα των τυχαιοποιημένων δειγμάτων που καλείται να κρίνει ο ακροατής, το πρωτόκολλο MUSHRA εισάγει κρυφές σταθερές ελέγχου για τη διασφάλιση της στατιστικής εγκυρότητας:

- Κρυφή Άγκυρα Αναφοράς (Hidden Reference): Ένα πανομοιότυπο αντίγραφο του αρχικού, μη συμπιεσμένου σήματος υψηλής πιστότητας. Χρησιμεύει για την επαλήθευση της ακουστικής οξύτητας και της αξιοπιστίας του ίδιου του αξιολογητή, ο οποίος οφείλει να του αποδώσει την ανώτατη βαθμολογία (100).
- Υποβαθμισμένη Άγκυρα (Anchor): Ένα σκόπιμα αλλοιωμένο δείγμα χαμηλής ποιότητας (για παράδειγμα, επεξεργασμένο με χαμηλοπερατό φίλτρο στα 3,5 kHz). Η άγκυρα αυτή θέτει το κατώτατο όριο της αντιληπτικής κλίμακας για το συγκεκριμένο τεστ.
- Έξοδοι των Μοντέλων: Τα ηχητικά αποσπάσματα που έχουν παραχθεί από τα υπό εξέταση παραγωγικά δίκτυα τεχνητής νοημοσύνης (π.χ. μοντέλα διάχυσης, αυτοπαλίνδρομα συστήματα ή μοντέλα ταύτισης ροής).

Οι αξιολογητές έχουν τη δυνατότητα να εναλλάσσουν την ακρόαση μεταξύ των δειγμάτων σε πραγματικό χρόνο και βαθμολογούν όλες τις συνθήκες παράλληλα επάνω σε μια συνεχή αριθμητική κλίμακα από το 0 έως το 100, η οποία είναι χωρισμένη στις κλασικές ποιοτικές κατηγορίες (Bad, Poor, Fair, Good, Excellent).

5.4 Εκπαίδευση μοντέλων και σύγκριση αποτελεσμάτων(FAD-CLAP)

Σε αυτή την υποενότητα αναλύεται η διαδικασία εκπαίδευσης, παραγωγής και ποσοτικής αξιολόγησης δύο διαφορετικών αρχιτεκτονικών τεχνητής νοημοσύνης για τη σύνθεση μουσικής disco. Η μεθοδολογική προσέγγιση διαρθρώνεται σε τρία διακριτά στάδια: αρχικά, παρουσιάζεται η αποδοτική ως προς τις παραμέτρους fine-tuning εκπαίδευση του μοντέλου βασισμένου σε Transformer (MusicGen) με τη χρήση ενός συνόλου δεδομένων 100 ακουστικών δειγμάτων. Ακολούθως, πραγματοποιείται συγκριτική μαζική παραγωγή (batch inference) νέων κομματιών διάρκειας 10sec, θέτοντας σε λειτουργία τόσο το προσαρμοσμένο μοντέλο Transformer όσο και μια αρχιτεκτονική λανθάνουσας διάχυσης (AudioLDM / Diffusion) υπό κοινές, αλγοριθμικά παραγόμενες κειμενικές οδηγίες (prompts). Η ενότητα ολοκληρώνεται με την αντικειμενική αποτίμηση των αποτελεσμάτων μέσω μετρικών, όπου εξετάζεται η νοηματική ευθυγράμμιση κειμένου-ήχου (CLAP) και η ακουστική πιστότητα των παραχθέντων δειγμάτων (FAD). Για τον υπολογισμό της ακουστικής απόκλισης, ως στατιστική βάση αναφοράς και μέτρο σύγκρισης χρησιμοποιούνται 100 αυθεντικά studio κομμάτια disco διάρκειας 30sec, προερχόμενα από το αναγνωρισμένο σύνολο δεδομένων GTZAN (διαθέσιμο μέσω της πλατφόρμας Hugging Face), προσφέροντας έτσι ένα ολοκληρωμένο και επιστημονικά ελεγχόμενο πλαίσιο αξιολόγησης των δύο τεχνολογικών προσεγγίσεων. Η υλοποίηση του κώδικα έχει πραγματοποιηθεί στην πλατφόρμα Google Collab.

5.4.1 Εκπαίδευση με βάση την αρχιτεκτονική Transformer

1) Εκπαίδευση Μοντέλου (Fine-Tuning)

Αυτή η εκτέλεση υλοποιεί μια αποδοτική ως προς τις παραμέτρους εκπαίδευση του μοντέλου MusicGen-small σε εξειδικευμένα δεδομένα ήχου διάρκειας 5 δευτερολέπτων (32 kHz). Για την εξοικονόμηση μνήμης VRAM, οι βασικοί μετασχηματιστές (transformers) παρέμειναν «παγωμένοι», ενώ οι ενημερώσεις των βαρών περιορίστηκαν αποκλειστικά στα επίπεδα κειμένου και εξόδου. Η βελτιστοποίηση έγινε με τον αλγόριθμο AdamW σε περιβάλλον αυτόματης μικτής ακρίβειας (AMP) για μέγιστη σταχύτητα και σταθερότητα. Η διαδικασία ολοκληρώθηκε μετά από 15 εποχές-επαναλήψεις(epochs).(ΠΑΡΑΡΤΗΜΑ Α - [Block 1](#))

2) Αυτοματοποιημένη Διαδικασία Παραγωγής (Inference)

Εδώ αναπτύσσεται μια αυτοματοποιημένη διαδικασία μαζικής παραγωγής ήχου (batch inference) για την αξιολόγηση του εκπαιδευμένου μοντέλου. Αλγοριθμικά έχουν συντεθεί 100 μοναδικές περιγραφές disco, συνδυάζοντας διαφορετικές εποχές, διαθέσεις, ρυθμούς και όργανα. Το μοντέλο, έχοντας ενσωματώσει τα νέα fine-tuned βάρη, παρήγαγε κομμάτια διάρκειας 10 δευτερολέπτων με τη στρατηγική δειγματοληψίας top-k. Επίσης παράγεται και ένα αρχείο manifest το οποίο λειτουργεί ως ένας κατάλογος μεταδεδομένων, ο οποίος καταγράφει αυτόματα το αναγνωριστικό, το μονοπάτι αποθήκευσης και την αντίστοιχη κειμενική περιγραφή κάθε παραχθέντος ήχου, με σκοπό να διασφαλιστεί η αυτοματοποιημένη και έγκυρη αντιστοίχιση των δειγμάτων κατά τη μετέπειτα φάση της ποσοτικής αξιολόγησης από τους αλγορίθμους CLAP και FAD.(ΠΑΡΑΡΤΗΜΑ Α - [Block 2](#))

5.4.2 Εκπαίδευση με βάση την αρχιτεκτονική Diffusion

Αυτή η εκτέλεση υλοποιεί μια διαδικασία μαζικής παραγωγής ήχου χρησιμοποιώντας το μοντέλο λανθάνουσας διάχυσης AudioLDM-small-v2. Χρησιμοποιώντας τις ίδιες ακριβώς 100 συνδυαστικές περιγραφές disco για λόγους αντικειμενικής σύγκρισης, ο αλγόριθμος εκτελεί μια αντίστροφη διαδικασία αποθορυβοποίησης 25 βημάτων χωρίς υπολογισμό

κλίσεων. Κάθε παραγόμενο κομμάτι έχει διάρκεια 10sec με εγγενή συχνότητα δειγματοληψίας 16 kHz. Η διαδικασία ολοκληρώνεται με την παραγωγή των αρχείων ήχου και τη δημιουργία ενός νέου, αυτόνομου αρχείου manifest, θέτοντας τις βάσεις για τη συγκριτική αξιολόγηση των δύο διαφορετικών αρχιτεκτονικών (Transformer έναντι Diffusion).(ΠΑΡΑΡΤΗΜΑ Α - [Block 3](#))

5.4.3 Σύγκριση μετρικών για κάθε αρχιτεκτονική

Με βάσει τις εκτελέσεις κώδικα για τον υπολογισμό των μετρικών CLAP και FAD για κάθε αρχιτεκτονική που χρησιμοποιήθηκε για τη δημιουργία 100 μουσικών κλιπ, τα αποτελέσματα δύο παρτίδων παρουσιάζουν χαοτικές διαφορές παρόλο που χρησιμοποιήθηκαν οι ίδιες ακριβώς 100 κειμενικές οδηγίες (prompts).(ΠΑΡΑΡΤΗΜΑ Α - [Block 4](#) - [Block 5](#) - [Block 6](#) - [Block 7](#))

Οι τιμές που καταγράφηκαν στα προηγούμενα βήματα αποδεικνύουν την καθολική υπεροχή της Transformer (MusicGen) αρχιτεκτονικής έναντι της Diffusion (AudioLDM) για το συγκεκριμένο είδος μουσικής (disco). Αξίζει να επισημανθεί ότι η σαφής αυτή ποιοτική διαφοροποίηση προέκυψε υπό τους αυστηρούς υπολογιστικούς περιορισμούς του πειράματος, τόσο ως προς τη διαθέσιμη μνήμη VRAM της GPU του Google Collab (γεγονός που επέβαλε τη στρατηγική μερικού παγώματος των επιπέδων του Transformer για τη διασφάλιση της σταθερότητας του συστήματος) όσο και ως προς τη σύντομη χρονική διάρκεια παραγωγής των ακουστικών κλιπ (10 δευτερόλεπτα). Τα αποτελέσματα φαίνονται στο παρακάτω πίνακα:

Μετρική Αξιολόγησης	Transformer (MusicGen-small Fine-Tuned)	Diffusion (AudioLDM-small- v2)	Ερμηνεία & Σημασία
Μέση Συνημίτονη Ομοιότητα CLAP	0.2956	-0.0021	Υψηλότερο = Καλύτερο. Δείχνει πόσο πιστά ακολουθεί το μοντέλο τις οδηγίες κειμένου (prompts).
Fréchet Audio Distance (FAD)	58.9986	177.3982	Χαμηλότερο = Καλύτερο. Δείχνει πόσο κοντά είναι ο ήχος ποιοτικά και στυλιστικά στην πραγματική disco (GTZAN).

Πίνακας 2: Πίνακας αξιολόγησης μετρικών FAD-CLAP

Συμπεράσματα από τη Σύγκριση:

- **Κατανόηση Κειμένου (CLAP):** Παρά το στενό παράθυρο εκπαίδευσης, το fine-tuned MusicGen πέτυχε σταθερά θετική συσχέτιση (0.2956), ενσωματώνοντας επιτυχώς δομικά στοιχεία του prompt (όπως strings και slap bass). Αντίθετα, το AudioLDM παρουσίασε ουσιαστικά μηδενική συσχέτιση (-0.0021), αδυνατώντας να αποδώσει τις κειμενικές λεπτομέρειες στον περιορισμένο χρόνο των 25 βημάτων αποθορυβοποίησης.
- **Ακουστική Πιστότητα & Στυλ (FAD):** Η απόσταση FAD του Transformer παρέμεινε σε χαμηλά επίπεδα (58.99), αποδεικνύοντας ότι η παραμετρικά αποδοτική μέθοδος fine-tuning (parameter-efficient) επέτρεψε στο μοντέλο να προσεγγίσει τη

στατιστική κατανομή της πραγματικής disco. Το σκορ 177.39 του Diffusion φανερώνει έντονη ακουστική απόκλιση από τα studio disco-δείγματα, επηρεασμένο πιθανώς από την FP16 συμπίεση και τον εγγενή θόρυβο της διαδικασίας διάχυσης σε σύντομα ηχητικά παράθυρα.

- **Προσωπική εκτίμηση:** Στο πραγματικό άκουσμα των 100 παραγόμενων κομματιών του μουσικού είδους disco, εκείνα που δημιουργήθηκαν υπό την αρχιτεκτονική Transformer ήταν πιο «ολοκληρωμένα», δηλαδή ακουγόντουσαν για παράδειγμα πραγματικά κύμβαλα και μελωδίες μπάσου και μουσική disco όπως αναμενόταν. Ενώ σε εκείνα υπό την αρχιτεκτονική Diffusion ακουγόταν ένας «χαλασμένος» ήχος με «φυσήματα» και με σαφώς δυσκολία αναγνώρισης του μουσικού είδους. Αυτή η προσωπική εκτίμηση αποτυπώνεται ξεκάθαρα και στα αποτελέσματα των παραπάνω μετρικών CLAP και FAD.

6. Συμπεράσματα, περιορισμοί και μελλοντικοί στόχοι

6.1 Συμπεράσματα

Στην παρούσα εργασία αναλύθηκαν συστηματικά τα αλγοριθμικά γρανάζια και τα μοντέλα μηχανικής μάθησης που υλοποιούν τη σύγχρονη παραγωγή ήχου και μουσικής από κείμενο. Η μετατροπή μιας αφηρημένης γλωσσικής περιγραφής σε μια πυκνή, συνεχή ακουστική κυματομορφή αποτελεί, χωρίς αμφιβολία, μία από τις πιο σύνθετες εργασίες διατροφικής χαρτογράφησης (cross-modal mapping) στη βαθιά μάθηση. Η ανάλυση έδειξε ότι η επιλογή της αρχιτεκτονικής επιβάλλει έναν αναπόφευκτο συμβιβασμό ανάμεσα στη μακρο-συνθετική δομή και τη μικρο-ακουστική πιστότητα.

Αυτό το διατροφικό ταξίδι ξεκινά από το front-end του κειμένου, όπου τα μεγάλα γλωσσικά μοντέλα στρώνουν το σημασιολογικό υπόβαθρο που απαιτείται για να αποκωδικοποιηθεί η ανθρώπινη πρόθεση. Ωστόσο, το πραγματικό μηχανικό φορτίο το σηκώνουν οι νευρωνικοί κωδικοποιητές ήχου. Αυτοί οι μηχανισμοί εφαρμόζουν Κβαντοποίηση Διανυσμάτων (VQ/RVQ) για να συμπιέσουν τα ακατέργαστα ηχητικά κύματα σε παράλληλα πλέγματα χαμηλής συχνότητας. Με αυτόν τον μετασχηματισμό, η χρονική διαστατικότητα των ηχητικών δεδομένων συρρικνώνεται, επιτρέποντας στα δίκτυα βαθιάς μάθησης να επεξεργάζονται ακουστικά σήματα χωρίς να εξαντλούν αμέσως τη μνήμη της GPU.

Αξιολογώντας τις βασικές γενετικές αρχιτεκτονικές, στην εργασία αυτή βρέθηκε ένα ξεκάθαρο χάσμα στα προφίλ απόδοσής τους:

- **Αυτοπαλίνδρομα Γλωσσικά Μοντέλα:** Χάρη στην αιτιώδη κάλυψη προσοχής (causal attention masking), αποδίδουν εξαιρετικά στη μοντελοποίηση αυστηρών χρονολογικών ιεραρχιών, ρυθμικών πλεγμάτων και της ακριβούς έναρξης των νοτών. Στον αντίποδα, υποφέρουν από απελπιστικά αργές, γραμμικές ταχύτητες συμπερασμού (πολυπλοκότητα $O(S)$) και είναι ευάλωτα στην προκατάληψη έκθεσης (exposure bias) όσο μεγαλώνει ο ορίζοντας παραγωγής.

- **Λανθάνοντα Μοντέλα Διάχυσης (Latent Diffusion Models):** Αυτές οι αρχιτεκτονικές προσπερνούν εντελώς το διαδοχικό χρονικό βήμα. Υπολογίζουν τις ενημερώσεις σε ολόκληρο τον χρονικό καμβά ταυτόχρονα, πετυχαίνοντας σταθερή πολυπλοκότητα $O(1)$. Αυτό το παράλληλο πλαίσιο προσφέρει κορυφαία ακουστική πιστότητα και οργανικές υφές, αλλά συχνά «χάνει» στον ρυθμό (ρυθμική μετατόπιση) και δυσκολεύεται να ελέγξει τη μακροδομή της σύνθεσης σε κομμάτια διάρκειας πολλών λεπτών.

6.2 Περιορισμοί

6.2.1 Το πρόβλημα της συνοχής σε μεγάλο εύρος: Γιατί τα μοντέλα δυσκολεύονται να διατηρήσουν ένα συνεπές μουσικό θέμα για αρκετά λεπτά

Η διατήρηση της συνοχής όταν παράγεται μουσική διάρκειας πολλών λεπτών αποτελεί τεράστια πρόκληση. Το πρόβλημα πηγάζει από την ακραία πυκνότητα των διακριτικών (tokens) του ψηφιακού ήχου και τους αυστηρούς υπολογιστικούς περιορισμούς των τυπικών αρχιτεκτονικών Transformer. Αντίθετα με ό,τι συμβαίνει στο κείμενο ή στις εικόνες, οι νευρωνικοί κωδικοποιητές ήχου υψηλής πιστότητας συμπιέζουν τα ακουστικά κύματα σε πυκνά, παράλληλα πλέγματα κωδικοβιβλίων. Σε έναν κωδικοποιητή που λειτουργεί με ρυθμό καρέ 50Hz σε τέσσερα επίπεδα Υπολειμματικής Διανυσματικής Κβαντοποίησης (RVQ)· αυτός ο μηχανισμός απαιτεί την παρακολούθηση 200 διακριτών στοιχείων ανά δευτερόλεπτο. Αν κλιμακώσουμε αυτή την απαίτηση σε ένα τυπικό μουσικό κομμάτι 4,75 λεπτών, το μήκος της ακολουθίας εκτινάσσεται στα 57.000 διακριτικά στοιχεία ($4,75 \cdot 60 \cdot 50 \cdot 4$). Καθώς ο μηχανισμός αυτοπροσοχής (self-attention) στους κλασικούς Transformers κλιμακώνεται τετραγωνικά ως προς το μήκος της ακολουθίας (πολυπλοκότητα $O(N^2)$), ο υπολογισμός ενός ενιαίου παγκόσμιου πίνακα προσοχής σε δεκάδες χιλιάδες συνεχόμενα tokens εξαντλεί αμέσως τη φυσική VRAM της GPU.

Αυτό το φράγμα υλικού αναγκάζει τα μοντέλα να εκπαιδεύονται σε περικομμένα τμήματα 10 έως 30 δευτερολέπτων. Έτσι, όταν το σύστημα καλείται κατά το inference να δημιουργήσει ολόκληρα κομμάτια χρησιμοποιώντας συρόμενα παράθυρα ή προσωρινή αποθήκευση ιστορικού, περιορίζεται αναγκαστικά στην τοπική συνοχή. Προβλέπει, δηλαδή, με επιτυχία την αμέσως επόμενη συγχορδία ή έναν σύντομο ήχο, αλλά παραμένει τυφλό απέναντι στον κυρίαρχο μακροδομικό χάρτη της σύνθεσης.

Αυτή η δομική τύφλωση γίνεται ακόμα πιο έντονη εξαιτίας της προκατάληψης έκθεσης (exposure bias) κατά τη φάση του συμπερασμού, αλλά και των σημασιολογικών περιορισμών του στατικού κειμένου (prompt). Κατά την αυτοπαλίνδρομη γένεση, το μοντέλο τροφοδοτεί το παράθυρο πλαισίου του με τις δικές του πιθανοτικές προβλέψεις, αντί να βασίζεται σε πραγματικά δεδομένα (ground-truth). Το αποτέλεσμα είναι ότι οποιοδήποτε απειροελάχιστο σφάλμα εισαχθεί σε ένα πρώιμο καρέ — μια ρυθμική καθυστέρηση μικροδευτερολέπτων ή μια μικρή αρμονική απόκλιση — εγκλωβίζεται μόνιμα στο πλαίσιο. Το λάθος αυτό μεγαλώνει εκθετικά με την πάροδο του χρόνου, αποσυντονίζει τη χρονογραμμή, καταστρέφει την ευθυγράμμιση με τον μετρονόμο και τελικά διαλύει τη δομή σε χαοτικά μοτίβα συχνοτήτων.

Την ίδια στιγμή, η ασυμμετρία των πληροφοριών δημιουργεί ένα σημείο απόλυτης συμφόρησης. Ένα τυπικό μήνυμα κειμένου αντιστοιχίζεται σε ένα ενιαίο, στατικό σημασιολογικό διάνυμα, το οποίο στέλνει ένα ομοιόμορφο, αμετάβλητο σήμα καθοδήγησης σε ολόκληρη τη χρονική σειρά των 5 λεπτών. Επειδή αυτό το διάνυμα δεν διαθέτει χρονικά ευθυγραμμισμένους δομικούς δείκτες για να σηματοδοτήσει αλλαγές (όπως το πέρασμα από το κουπλέ στο ρεφρέν), ο γενετικός πυρήνας καταφεύγει σε τυχαίες στατιστικές μεταβάσεις. Αυτό συχνά ομογενοποιεί το κομμάτι σε μια επαναλαμβανόμενη λούπα, στερώντας του τη δυναμική εξέλιξη που χαρακτηρίζει τις ανθρώπινες συνθέσεις

6.2.2 Περιορισμοί Υπολογισμού και Υλικού

Η διατροφική σύνθεση κειμένου σε ήχο δεν περιορίζεται τόσο από την ακατέργαστη υπολογιστική ταχύτητα των επεξεργαστών, όσο από τη μνήμη του υλικού. Το αποτέλεσμα

Διπλωματική Εργασία

είναι διπλό: ακραία εξάντληση της VRAM κατά την εκπαίδευση και σοβαρός κορεσμός του εύρους ζώνης μνήμης κατά το συμπερασμό (inference). Η αιτία για το OOM (Out-of-Memory) στην εκπαίδευση είναι η μαθηματική διατύπωση του πίνακα προσοχής Scaled Dot-Product (Ενότητα 2.3.2). Επειδή ο πίνακας αυτός κλιμακώνεται τετραγωνικά ως προς το μήκος της ηχητικής ακολουθίας, προκαλεί μια πολυπλοκότητα χώρου $O(S^2)$. Η αποθήκευση του ενδιάμεσου τανυστή βάρους προσοχής με σχήμα $[B, H, S, S]$ απαιτεί τεράστια επιβάρυνση. Η GPU είναι αναγκασμένη να διατηρεί αυτές τις ακριβείς καταστάσεις ενεργοποίησης στη μνήμη υψηλού εύρους ζώνης (HBM) σε όλα τα επίπεδα, ώστε να μπορεί να εκτελέσει την οπισθοδιάδοση (backpropagation).

- **Το παράδειγμα των 5 λεπτών:** Για έναν τυπικό ρυθμό καρέ ήχου 50Hz, ένα κομμάτι 5 λεπτών μεταφράζεται σε 15.000 καρέ. Αυτό δημιουργεί $2,25 \cdot 10^8$ στοιχεία, καταναλώνοντας περίπου 14,4 Gigabytes μνήμης ενεργοποίησης ανά επίπεδο με ακρίβεια FP32 (Κινητή Υποδιαστολή 32-bit).
- **Το αδιέξοδο:** Όταν αυτή η απαίτηση πολλαπλασιάζεται σε έναν τυπικό κορμό Transformer 24 έως 48 επιπέδων, το αποτύπωμα ξεπερνά γρήγορα τη φυσική χωρητικότητα ακόμα και των κορυφαίων GPUs των 80GB. Αυτός είναι ο λόγος που τα μοντέλα μετατροπής κειμένου σε μουσική περιορίζονταν ιστορικά σε σύντομα ηχητικά θραύσματα κατά την εκπαίδευση.

Κατά τον χρόνο εκτέλεσης (runtime), το πρόβλημα μετατοπίζεται στην προσωρινή μνήμη Key-Value (KV Cache), μετατρέποντας τη διαδικασία σε μια εργασία που περιορίζεται αυστηρά από το εύρος ζώνης της μνήμης (memory-bandwidth bound). Για να αποφευχθεί ο περιττός επαναυπολογισμός των τιμών προσοχής για τα ιστορικά tokens σε κάθε βήμα, τα αυτοπαλίνδρομα μοντέλα αποθηκεύουν τα διανύσματα Key (K) και Value (V) στη VRAM. Αυτό δημιουργεί ένα αποτύπωμα που κλιμακώνεται γραμμικά ($O(S)$) σε σχέση με το μέγεθος της παρτίδας, το βάθος της στρώσης, το μήκος της ακολουθίας και την κρυφή διάσταση. Στη μουσική μακράς φόρμας, η αποθήκευση αυτών των διανυσμάτων υψηλής διάστασης σε παράλληλα επίπεδα βιβλίου κωδικών καταναλώνει δεκάδες gigabytes. Οι επεξεργαστικοί

Διπλωματική Εργασία

πυρήνες της GPU μένουν άπραγοι (idle), περιμένοντας την ανάκτηση αυτών των μαζικών τανυστών από την αργή HBM στην εξαιρετικά γρήγορη SRAM on-chip.

Για να προσπεραστούν οι περιορισμοί αυτοί, τα σύγχρονα συστήματα επιστρατεύουν δύο βασικές τεχνικές:

- **FlashAttention:** Εφαρμόζει πλακίδια πινάκων με επίγνωση των λειτουργιών I/O (IO-aware tiling) για να διαιρέσει τους πίνακες προσοχής σε μικρότερα, εντοπισμένα μπλοκ. Εκτελεί σταδιακά την κανονικοποίηση softmax μέσα στην εξαιρετικά γρήγορη SRAM, μειώνοντας την πολυπλοκότητα χώρου της προσοχής από τετραγωνική ($O(S^2)$) σε αυστηρά γραμμική ($O(S)$).
- **Grouped-Query Attention (GQA):** Ομαδοποιεί γειτονικές κεφαλές ερωτημάτων (query heads) ώστε να μοιράζονται μια ενοποιημένη διαδρομή Κλειδιού-Τιμής (KV). Συμπιέζει την κάθετη διάσταση του πίνακα της KV cache κατά παράγοντες 4 ή 8, μειώνοντας άμεσα τον κορεσμό του διαύλου μνήμης και επιτρέποντας υψηλή απόδοση σε πραγματικό χρόνο.

6.3 Προοπτικές για τις Μελλοντικές Κατευθύνσεις της Έρευνας

Αν και το πολυσταδιακό υβριδικό παράδειγμα σταθεροποιεί τη βασική γραμμή παραγωγής μουσικής από κείμενο, αρκετοί μηχανικοί περιορισμοί παραμένουν ανοιχτοί, χαράσσοντας την επόμενη πορεία της ακαδημαϊκής έρευνας για την παραγωγή μουσικής από κείμενο.

1. Μετάβαση στη Γραμμική Χρονική Δυναμική μέσω Μοντελοποίησης Χώρου Καταστάσεων (SSM)

Για να ξεπεραστεί οριστικά το τείχος του τετραγωνικού κόστους ενεργοποιήσεων ($O(S^2)$), το οποίο επιβάλλουν οι μηχανισμοί αυτοπροσοχής πολλαπλών κεφαλών (multi-head self-attention) σε μεγάλες ηχητικές ακολουθίες, οι μελλοντικές αρχιτεκτονικές αντικαθιστούν ενεργά τα τυπικά μοντέλα Transformer. Στη θέση τους αναδύονται τα Μοντέλα Χώρου

Καταστάσεων (State Space Models - SSM), όπως το Mamba-2. Τα SSM διατηρούν μια εσωτερική επαναλαμβανόμενη κατάσταση σταθερού μεγέθους. Κατά την εκπαίδευση, επεξεργάζονται τα tokens με γραμμικές, χρονικά αμετάβλητες συνελίξεις, ενώ κατά το inference επιστρέφουν σε μια αποτελεσματική εκτέλεση τύπου RNN. Με αυτόν τον τρόπο, ξεκλειδώνουν τη δυνατότητα επεξεργασίας ολοκληρωμένων ενορχηστρώσεων πολλαπλών στελεχών (multi-stem) από την αρχή, παραμένοντας εντός πολύ περιορισμένων ορίων μνήμης GPU.

2. Λανθάνουσα Ροή κάτω των 150ms για Διαδραστική Δημιουργία σε Πραγματικό Χρόνο

Οι τρέχουσες μηχανές συμπερασμού απαιτούν τη δημιουργία ενός ολόκληρου μπλοκ μουσικής προτού ξεκινήσει η αναπαραγωγή. Η επόμενη γενιά διαδραστικών διεπαφών ήχου —όπως οι ζωντανές συνεδρίες jam με τη βοήθεια AI ή οι μηχανές δυναμικής προσαρμογής soundtrack σε βιντεοπαιχνίδια— απαιτεί βρόχους ροής (streaming) με εξαιρετικά χαμηλή καθυστέρηση. Αυτή η ερευνητική κατεύθυνση βασίζεται στην ανάπτυξη αιτιακών δικτύων (causal networks) με επίγνωση πλαισίου και πεπερασμένης βαθμωτής κβαντοποίησης (finite scalar quantization). Διαιρώντας το πέρασμα αποθορυβοποίησης της λανθάνουσας διάχυσης σε διαδοχικά, επικαλυπτόμενα χρονικά μπλοκ, τα συστήματα μπορούν να πετύχουν καθυστέρηση ροής κάτω από 180°C ή για την ακρίβεια κάτω από 150ms κατά τον χρόνο εκτέλεσης. Έτσι, αποδίδουν ήχο υψηλής πιστότητας συνεχώς, καθώς ο χρήστης τροποποιεί τα prompts κειμένου σε πραγματικό χρόνο.

3. Ευθυγράμμιση Ανθρώπινων Προτιμήσεων μέσω Άμεσης Βελτιστοποίησης (DPO)

Οι παραδοσιακές συναρτήσεις απώλειας (loss functions) αναγκάζουν τα μοντέλα να μιμούνται τυφλά την κατανομή των δεδομένων εκπαίδευσης, χωρίς να συνυπολογίζουν την καλλιτεχνική ποιότητα του ήχου. Για να γεφυρωθεί το χάσμα μεταξύ της στατιστικής σύγκλισης και της ανθρώπινης μουσικότητας, οι μελλοντικές εργασίες ενσωματώνουν την

offline και online Άμεση Βελτιστοποίηση Προτιμήσεων (Direct Preference Optimization - DPO) απευθείας στη διαδρομή επιλογής της ακολουθίας ήχου. Συλλέγοντας ζεύγη δεδομένων με ανθρώπινες προτιμήσεις (όπως το MusicPrefs) και υπολογίζοντας μια απώλεια αντίθεσης (contrastive loss) πάνω σε συγκεκριμένους άξονες, οι γενετικοί πυρήνες μαθαίνουν να απορρίπτουν ακουστικά σφάλματα. Οι βασικοί αυτοί άξονες περιλαμβάνουν την καθαρή ευθυγράμμιση των στίχων, την αρμονική συνέπεια, την απόλυτη καθαρότητα της μίξης. Με αυτή τη μέθοδο, τα μοντέλα βελτιστοποιούνται ώστε να δίνουν προτεραιότητα στα σύνθετα δομικά μοτίβα που το ανθρώπινο αυτί αντιλαμβάνεται ως ολοκληρωμένη και συναισθηματικά φορτισμένη μουσική.

Βιβλιογραφία

- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1), 147–169. [https://doi.org/10.1016/S0364-0213\(85\)80012-4](https://doi.org/10.1016/S0364-0213(85)80012-4)
- Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. (2023). MusicLM: Generating Music From Text. *ArXiv:2301.11325 [Cs, Eess]*. <https://arxiv.org/abs/2301.11325>
- Alexandre Défossez, Copet, J., Synnaeve, G., & Adi, Y. (2022). High Fidelity Neural Audio Compression. *ArXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2210.13438>
- Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157–166. <https://doi.org/10.1109/72.279181>
- Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Olivier Pietquin, Sharifi, M., Olivier Teboul, Grangier, D., Tagliasacchi, M., & Zeghidour, N. (2022). AudioLM: a Language Modeling Approach to Audio Generation. *ArXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2209.03143>
- brweb. (2021). BS.1534: Method for the subjective assessment of intermediate quality level of audio systems. Itu.int. <https://www.itu.int/rec/R-REC-BS.1534-1-200301-S/en>

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A.

(2023, June 8). *Simple and Controllable Music Generation*. ArXiv.org.

<https://doi.org/10.48550/arXiv.2306.05284>

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North, 1*, 4171–4186. <https://doi.org/10.18653/v1/n19-1423>

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A Generative Model for Music. *ArXiv:2005.00341 [Cs, Eess, Stat]*.

<https://arxiv.org/abs/2005.00341>

Elizalde, B., Deshmukh, S., Ismail, M. A., & Wang, H. (2022, June 9). *CLAP: Learning Audio Concepts From Natural Language Supervision*. ArXiv.org.

<https://doi.org/10.48550/arXiv.2206.04769>

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, 14(2), 179–211.

https://doi.org/10.1207/s15516709cog1402_1

Evans, Z., Parker, J. D., Carr, C., Zukowski, Z., Taylor, J., & Pons, J. (2024). *Stable Audio Open*. ArXiv.org. <https://arxiv.org/abs/2407.14358>

Fan, A., Lewis, M., & Dauphin, Y. (2018). *Hierarchical Neural Story Generation*.

ArXiv.org. <https://arxiv.org/abs/1805.04833>

Fourier, J.B. (1822) *The Analytical Theory of Heat-English*. The University Press, 408. -

References - Scientific Research Publishing. (2017). Scirp.org.

<https://www.scirp.org/reference/referencespapers?referenceid=2158142>

Gabor, D. (n.d.). *THEORY OF COMMUNICATION**. <https://jcsphysics.net/lit/gabor1946.pdf>

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual

Prediction with LSTM. *Neural Computation*, 12(10), 2451–2471.

<https://doi.org/10.1162/089976600300015015>

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Deeplearningbook.org;

MIT Press. <https://www.deeplearningbook.org/>

Griffin, D., & Jae Lim. (1984). Signal estimation from modified short-time Fourier transform.

IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2), 236–243.

<https://doi.org/10.1109/TASSP.1984.1164317>

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z.,

Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for

Speech Recognition. *ArXiv:2005.08100 [Cs, Eess]*. <https://arxiv.org/abs/2005.08100>

Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier

transform. *Proceedings of the IEEE*, 66(1), 51–83.

<https://doi.org/10.1109/proc.1978.10837>

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models.

ArXiv:2006.11239 [Cs, Stat]. <https://arxiv.org/abs/2006.11239>

- Ho, J., & Salimans, T. (2022). Classifier-Free Diffusion Guidance. *ArXiv:2207.12598 [Cs]*.
<https://arxiv.org/abs/2207.12598>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The Curious Case of Neural Text Degeneration. *ArXiv:1904.09751 [Cs]*. <https://arxiv.org/abs/1904.09751>
- Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., & Carreira, J. (2021). Perceiver: General Perception with Iterative Attention. *ArXiv:2103.03206 [Cs, Eess]*.
<https://arxiv.org/abs/2103.03206>
- Kilgour, K., Zuluaga, M., Roblek, D., & Sharifi, M. (2019, January 17). *Fr'echet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms*. ArXiv.org.
<https://doi.org/10.48550/arXiv.1812.08466>
- Kingma, D. P., & Ba, J. (2014, December 22). *Adam: A Method for Stochastic Optimization*. ArXiv. <https://arxiv.org/abs/1412.6980>
- Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *ArXiv:2010.05646 [Cs, Eess]*.
<https://arxiv.org/abs/2010.05646>
- Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *ArXiv:1912.10211 [Cs, Eess]*. <https://arxiv.org/abs/1912.10211>

- Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., & Adi, Y. (2022). AudioGen: Textually Guided Audio Generation. *ArXiv:2209.15352 [Cs, Eess]*. <https://arxiv.org/abs/2209.15352>
- Kudo, T. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. *ArXiv:1804.10959 [Cs]*. <https://arxiv.org/abs/1804.10959>
- Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *ArXiv:1808.06226 [Cs]*. <https://arxiv.org/abs/1808.06226>
- Lipman, Y., Ricky, Ben-Hamu, H., Nickel, M., & Le, M. (2023). *Flow Matching for Generative Modeling*. ArXiv.org. <https://arxiv.org/abs/2210.02747>
- Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., & Plumbley, M. D. (2023). AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *ArXiv:2301.12503 [Cs, Eess]*. <https://arxiv.org/abs/2301.12503>
- Loshchilov, I., & Hutter, F. (2017). Decoupled Weight Decay Regularization. *Arxiv.org*. <https://arxiv.org/abs/1711.05101>
- Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., & Zhu, J. (2022, October 13). *DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps*. ArXiv.org. <https://doi.org/10.48550/arXiv.2206.00927>

- Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., & Zhu, J. (2023, May 6). *DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models*. ArXiv.org. <https://doi.org/10.48550/arXiv.2211.01095>
- Lu, Y., Li, Z., He, D., Sun, Z., Dong, B., Qin, T., Wang, L., & Liu, T.-Y. (2019, June 6). *Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View*. ArXiv.org. <https://doi.org/10.48550/arXiv.1906.02762>
- Mentzer, F., Minnen, D., Agustsson, E., & Tschannen, M. (2023). *Finite Scalar Quantization: VQ-VAE Made Simple*. ArXiv.org. <https://arxiv.org/abs/2309.15505>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, September 7). *Efficient Estimation of Word Representations in Vector Space*. ArXiv.org. <https://arxiv.org/abs/1301.3781>
- Oord, A. van den, Vinyals, O., & Kavukcuoglu, K. (2018). Neural Discrete Representation Learning. *ArXiv:1711.00937 [Cs]*. <https://arxiv.org/abs/1711.00937>
- Oppenheim, A.V. and Schaffer, R.W. (2010) *Discrete-Time Signal Processing*. Pearson Higher Education, Inc., Upper Saddle River. - References - Scientific Research Publishing. (2024). Scirp.org. <https://www.scirp.org/reference/referencespapers?referenceid=3718017>
- P.800 :Methods for subjective determination of transmission quality. (n.d.). Wwww.itu.int. <https://www.itu.int/rec/T-REC-P.800-199608-I>
- Peebles, W., & Xie, S. (2023, December 19). *Scalable Diffusion Models with Transformers*. ArXiv.org. <https://arxiv.org/abs/2212.09748>

- Pennington, J., Socher, R., & Manning, C. (2014). *GloVe: Global Vectors for Word Representation* (pp. 1532–1543). Association for Computational Linguistics.
<https://aclanthology.org/D14-1162.pdf>
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. ArXiv.org.
<https://arxiv.org/abs/1802.05365>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *ArXiv:2103.00020 [Cs]*.
<https://arxiv.org/abs/2103.00020>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. ArXiv.org. <https://arxiv.org/abs/1910.10683>
- Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3), 400–407. <https://doi.org/10.1214/aoms/1177729586>
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *ArXiv:2112.10752 [Cs]*.
<https://arxiv.org/abs/2112.10752>

- Ronneberger, O., Fischer, P., & Brox, T. (2015, May 18). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. ArXiv.org. <https://arxiv.org/abs/1505.04597>
- S. Stevens, J. Stanley, J. Volkman and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *Journal of the Acoustical Society of America*, Vol. 8, No. 3, 1937, pp. 185-190. - References - Scientific Research Publishing. (2026). Scirp.org. <https://www.scirp.org/reference/referencespapers?referenceid=606529>
- Schuster, M., & Nakajima, K. (2012, March 1). *Japanese and Korean voice search*. IEEE Xplore. <https://doi.org/10.1109/ICASSP.2012.6289079>
- Sennrich, R., Haddow, B., & Birch, A. (2016). *Neural Machine Translation of Rare Words with Subword Units*. <https://arxiv.org/pdf/1508.07909>
- Song, J., Meng, C., & Ermon, S. (2022). Denoising Diffusion Implicit Models. *ArXiv:2010.02502 [Cs]*. <https://arxiv.org/abs/2010.02502>
- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *WaveNet: A Generative Model for Raw Audio*. ArXiv.org. <https://arxiv.org/abs/1609.03499>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention Is All You Need*. Cornell University. <https://arxiv.org/abs/1706.03762>
- Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, Rif A.

(2017). *Tacotron: Towards End-to-End Speech Synthesis*. ArXiv.org.

<https://arxiv.org/abs/1703.10135>

Williams, R. J., & Zipser, D. (1989). A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2), 270–280.

<https://doi.org/10.1162/neco.1989.1.2.270>

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Shlomo Dubnov. (2023).

Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. *ArXiv (Cornell University)*.

<https://doi.org/10.48550/arxiv.2211.06687>

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., & Tagliasacchi, M. (2021). SoundStream: An End-to-End Neural Audio Codec. *Arxiv.org*.

<https://doi.org/10.48550/arXiv.2107.03312>

Zwicker, E. and Fastl, H. (1999) *Psychoacoustics Facts and Models*. 3rd Edition, Springer, Heidelberg. - References - Scientific Research Publishing. (2026). Scirp.org.

<https://www.scirp.org/reference/referencespapers?referenceid=1152082>

Παράρτημα Α: Κώδικες Υλοποίησης Εκπαίδευσης Μοντέλων

Block 1

```
import gc
import os
import json
import torch
import torchaudio
from torch.optim import AdamW
from torch.utils.data import Dataset, DataLoader
from audiocraft.models import MusicGen
from tqdm import tqdm

# Forcing evict any residual tensor allocations to maximize VRAM
headroom
gc.collect()
torch.cuda.empty_cache()

# =====
# 1. HARDWARE-OPTIMIZED DATASET
# =====
class DiscoDataset(Dataset):
    def __init__(self, jsonl_path, target_sr=32000, duration=5):
        self.data = []
        self.target_sr = target_sr
        self.max_samples = target_sr * duration

        with open(jsonl_path, "r", encoding="utf-8") as f:
            for line in f:
                if line.strip():
                    self.data.append(json.loads(line))

    def __len__(self):
        return len(self.data)

    def __getitem__(self, idx):
        entry = self.data[idx]
        audio_path = entry["path"]
        description = entry["description"]

        try:
            wav, sr = torchaudio.load(audio_path)
            if wav.shape[0] > 1:
                wav = wav.mean(dim=0, keepdim=True)

            if sr != self.target_sr:
                resampler = torchaudio.transforms.Resample(sr,
self.target_sr)
                wav = resampler(wav)
```

```
        if wav.shape[1] > self.max_samples:
            wav = wav[:, :self.max_samples]
        elif wav.shape[1] < self.max_samples:
            pad_amount = self.max_samples - wav.shape[1]
            wav = torch.nn.functional.pad(wav, (0, pad_amount))

        return wav.squeeze(0), description
    except Exception as e:
        return torch.zeros(self.max_samples), description

# =====
# 2. PARAMETER-EFFICIENT TRAINING LOOP
# =====
def run_gpu_training():
    device = "cuda" if torch.cuda.is_available() else "cpu"
    print(f"? Execution Engine Initialized on: {device.upper()}")

    print("? Downloading pre-trained MusicGen weights...")
    model = MusicGen.get_pretrained('facebook/musicgen-small')

    # Keeping master language model parameters in stable Float32 space
    lm = model.lm.to(device).float()
    codec = model.compression_model.to(device)

    lm.train()
    codec.eval()

    # HARDWARE OPTIMIZATION: Freezing the bulk of the transformer
    layers
    print("? Freezing base transformer layers to secure VRAM safety
    margins...")
    for param in lm.parameters():
        param.requires_grad = False

    # Unfreezing only the final layer blocks and text conditioning
    adapters
    for param in lm.condition_provider.parameters():
        param.requires_grad = True
    if hasattr(lm, 'out'):
        for param in lm.out.parameters():
            param.requires_grad = True

    manifest_path = "/content/drive/MyDrive/audiocraft-
    project/clean_data.jsonl"
    dataset = DiscoDataset(manifest_path, target_sr=model.sample_rate,
    duration=5)
    dataloader = DataLoader(dataset, batch_size=1, shuffle=True)

    # Optimizer acting on active gradients, dramatically lowering
    overhead
```

```
optimizer = AdamW(filter(lambda p: p.requires_grad,
lm.parameters()), lr=5e-5, weight_decay=0.01)
scaler = torch.amp.GradScaler('cuda')

gc.collect()
torch.cuda.empty_cache()

print(f"? Dataset loaded successfully! Processing 100 tracks in
cloud memory.")
print("--- Starting True Gradient Optimization ---")

# RUNTIME CONFIGURATION
epochs = 15

for epoch in range(1, epochs + 1):
    epoch_loss = 0.0
    progress_bar = tqdm(dataloader, desc=f"Epoch {epoch}/{epochs}")

    for wavs, descriptions in progress_bar:
        wavs = wavs.unsqueeze(1).to(device).float()
        optimizer.zero_grad()

        with torch.no_grad():
            audio_tokens, scale = codec.encode(wavs)
            attributes, _ =
model._prepare_tokens_and_attributes(descriptions, None)

        with torch.amp.autocast('cuda'):
            output = lm.compute_predictions(codes=audio_tokens,
conditions=attributes)
            logits = output.logits
            mask = output.mask

            logits = torch.nan_to_num(logits, nan=0.0, posinf=1e4,
neginf=-1e4)

            loss = torch.nn.functional.cross_entropy(
                logits.reshape(-1, logits.shape[-1]),
                audio_tokens.reshape(-1),
                reduction='none'
            )
            loss = (loss * mask.reshape(-1)).sum() / max(mask.sum(),
1)

        scaler.scale(loss).backward()
        scaler.unscale_(optimizer)
        torch.nn.utils.clip_grad_norm_(filter(lambda p:
p.requires_grad, lm.parameters()), max_norm=1.0)

        scaler.step(optimizer)
        scaler.update()
```

```
epoch_loss += loss.item()
progress_bar.set_postfix({"Loss": f"{loss.item():.4f}"})

print(f"✔ Epoch {epoch} Completed! Average cross-entropy loss:
{epoch_loss / len(dataloader):.4f}")

# =====
# 3. SECURING THE FINE-TUNED WEIGHTS TO GDRIVE
# =====
output_dir = "/content/drive/MyDrive/audiocraft-project/models"
os.makedirs(output_dir, exist_ok=True)

# Saving file name:
save_path = os.path.join(output_dir,
"gpu_disco_transformer_weights1.th")

torch.save(lm.state_dict(), save_path)
print(f"\n? SUCCESS! Fine-tuned weights permanently secured to
Google Drive at:\n{save_path}")

run_gpu_training()
```

Block 2

```
import os
import json
import gc
import torch
import torchaudio
from audiocraft.models import MusicGen
from tqdm import tqdm

# Pre-flight hardware flushing to clear out memory from the training
phase
gc.collect()
torch.cuda.empty_cache()

device = "cuda" if torch.cuda.is_available() else "cpu"
print(f"? Evaluation Pipeline Active on: {device.upper()}")

# =====
# 1. ALGORITHMIC DISCO PROMPT GENERATOR
# =====
eras = ["1970s", "retro Studio 54", "vintage", "classic", "uptempo"]
moods = ["vibrant and energetic", "smooth and soulful", "fast-paced",
"grooves-heavy", "late-night"]
```

```
grooves = ["driving four-on-the-floor kick drum pattern", "syncopated  
open hi-hat rhythm", "heavy bass guitar pocket", "funky rhythmic  
backbeat"]  
leads = ["swirling string ensembles", "bright brass section stabs",  
"slap bass accents", "clean rhythm guitar chucks", "analog synthesizer  
leads"]  
  
eval_prompts = []  
for era in eras:  
    for mood in moods:  
        for groove in grooves:  
            for lead in leads:  
                prompt = f"A {mood} {era} disco track featuring a  
{groove} and {lead}."  
                if prompt not in eval_prompts:  
                    eval_prompts.append(prompt)  
  
eval_prompts = eval_prompts[:100]  
print(f"✦ Successfully generated {len(eval_prompts)} scientifically  
varied evaluation prompts.")  
  
# =====  
# 2. PIPELINE SYSTEM INITIALIZATION  
# =====  
print("? Loading baseline framework configuration...")  
model = MusicGen.get_pretrained('facebook/musicgen-small',  
device=device)  
  
weights_path = "/content/drive/MyDrive/audiocraft-  
project/models/gpu_disco_transformer_weights1.th"  
  
if not os.path.exists(weights_path):  
    raise FileNotFoundError(f"✗ Critical Error: Could not locate custom  
weights at {weights_path}")  
  
print("? Injecting newly fine-tuned transformer weights...")  
state_dict = torch.load(weights_path, map_location=device)  
model.lm.load_state_dict(state_dict)  
  
model.lm.to(device)  
model.compression_model.to(device)  
model.lm.float()  
  
model.set_generation_params(use_sampling=True, top_k=250, duration=10)  
  
# Directing output directory to folder  
output_dir = "/content/drive/MyDrive/audiocraft-  
project/evaluation_batch1"  
os.makedirs(output_dir, exist_ok=True)  
manifest_save_path = os.path.join(output_dir,  
"validation_manifest.jsonl")
```

```
if os.path.exists(manifest_save_path):
    os.remove(manifest_save_path)

# =====
# 3. AUTOMATED BATCH LOOP EXECUTION
# =====
print("\n--- Starting Upgraded 100-Song Automated Synthesis ---")

for idx, prompt in enumerate(tqdm(eval_prompts, desc="Generating
Batch")):
    track_id = f"eval_track_{idx+1:03d}"
    filename = f"{track_id}.wav"
    full_audio_save_path = os.path.join(output_dir, filename)

    with torch.no_grad():
        output_audio = model.generate([prompt], progress=False)

        output_audio = output_audio.detach().cpu().squeeze(0)
        torchaudio.save(full_audio_save_path, output_audio,
model.sample_rate)

        record = {"id": track_id, "path": full_audio_save_path,
"description": prompt}
        with open(manifest_save_path, "a", encoding="utf-8") as f:
            f.write(json.dumps(record, ensure_ascii=False) + "\n")

        del output_audio
        torch.cuda.empty_cache()

print(f"\n? SUCCESS! Generated {len(eval_prompts)} evaluation tracks in
Google Drive at:\n{output_dir}")
print(f"? Validation manifest file successfully rebuilt
at:\n{manifest_save_path}")
```

Block 3

```
import os
import json
import gc
import torch
import torchaudio
from diffusers import AudioLDMPipeline
from tqdm import tqdm

# Pre-flight hardware flushing to clear out GPU memory channels
gc.collect()
torch.cuda.empty_cache()
```

```
device = "cuda" if torch.cuda.is_available() else "cpu"
print(f"? Diffusion Pipeline Active on: {device.upper()}")

# =====
# 1. IDENTICAL ALGORITHMIC PROMPT GENERATOR
# =====
# Keeping these variables exactly identical to ensure a scientifically
controlled test
eras = ["1970s", "retro Studio 54", "vintage", "classic", "uptempo"]
moods = ["vibrant and energetic", "smooth and soulful", "fast-paced",
"grooves-heavy", "late-night"]
grooves = ["driving four-on-the-floor kick drum pattern", "syncopated
open hi-hat rhythm", "heavy bass guitar pocket", "funky rhythmic
backbeat"]
leads = ["swirling string ensembles", "bright brass section stabs",
"slap bass accents", "clean rhythm guitar chucks", "analog synthesizer
leads"]

eval_prompts = []
for era in eras:
    for mood in moods:
        for groove in grooves:
            for lead in leads:
                prompt = f"A {mood} {era} disco track featuring a
{groove} and {lead}."
                if prompt not in eval_prompts:
                    eval_prompts.append(prompt)

eval_prompts = eval_prompts[:100]
print(f"?* Successfully generated {len(eval_prompts)} identical
evaluation prompts.")

# =====
# 2. DIFFUSION MODEL INITIALIZATION
# =====
print("? Fetching pre-trained AudioLDM Diffusion weights...")
# Loading the small optimized version in FP16 precision to maximize
processing speed
model_id = "cvssp/audioldm-s-full-v2"
pipe = AudioLDMPipeline.from_pretrained(model_id,
torch_dtype=torch.float16)
pipe = pipe.to(device)

# Configuring the target directories to keep data safe and separate
output_dir = "/content/drive/MyDrive/audiocraft-
project/evaluation_batch_diffusion"
os.makedirs(output_dir, exist_ok=True)
manifest_save_path = os.path.join(output_dir,
"validation_manifest_diffusion.jsonl")

if os.path.exists(manifest_save_path):
    Διπλωματική Εργασία
```

```
os.remove(manifest_save_path)

# =====
# 3. AUTOMATED DIFFUSION GENERATION LOOP
# =====
print("\n--- Starting 100-Song Automated Diffusion Generation ---")

for idx, prompt in enumerate(tqdm(eval_prompts, desc="Diffusion
Generation")):
    track_id = f"diff_track_{idx+1:03d}" # Formatted as diff_track_001
    to 100
    filename = f"{track_id}.wav"
    full_audio_save_path = os.path.join(output_dir, filename)

    # Running the continuous reverse-diffusion denoising process
    with torch.no_grad():
        # audio_length_in_s=10.0 sets generation duration
        # num_inference_steps=25 handles the denoising steps efficiently
        audio_output = pipe(prompt, audio_length_in_s=10.0,
num_inference_steps=25).audios[0]

    # Converting the numpy array output back into a tensor for stable
    saving
    audio_tensor = torch.tensor(audio_output).unsqueeze(0)

    # AudioLDM naturally outputs at a crisp 16000Hz sampling rate
    torchaudio.save(full_audio_save_path, audio_tensor,
sample_rate=16000)

    # Appending data record to the manifest file
    record = {"id": track_id, "path": full_audio_save_path,
"description": prompt}
    with open(manifest_save_path, "a", encoding="utf-8") as f:
        f.write(json.dumps(record, ensure_ascii=False) + "\n")

    # Cleaning cache structures inside the loop
    del audio_output, audio_tensor
    torch.cuda.empty_cache()

print(f"\n? SUCCESS! Generated {len(eval_prompts)} diffusion tracks in
Google Drive at:\n{output_dir}")
print(f"? Diffusion manifest file successfully built
at:\n{manifest_save_path}")
```

Block 4 : (CLAP για Transformer)

```
import os
import json
import torch
import torchaudio
import torch.nn.functional as F
from transformers import ClapModel, ClapProcessor
from tqdm import tqdm

# Clearing active hardware cache to prevent out-of-memory (OOM) errors
torch.cuda.empty_cache()

device = "cuda" if torch.cuda.is_available() else "cpu"
print(f"? Initializing CLAP Evaluation Engine on: {device.upper()}")

print("? Fetching pre-trained LAION-CLAP embedding models...")
model_name = "laion/clap-htsat-fused"
processor = ClapProcessor.from_pretrained(model_name)
clap_model = ClapModel.from_pretrained(model_name).to(device)
clap_model.eval() # Setting model to evaluation mode (disables dropout,
batchnorm updates)

# Creating batch 1 directory:
manifest_path = "/content/drive/MyDrive/audiocraft-
project/evaluation_batch1/validation_manifest.jsonl"

if not os.path.exists(manifest_path):
    raise FileNotFoundError(f"✗ Cannot find validation_manifest.jsonl
at {manifest_path}")

print("? Validation manifest located. Initializing alignment scan...")

scores = []
with open(manifest_path, "r", encoding="utf-8") as f:
    lines = [line.strip() for line in f if line.strip()]

print(f"? Processing {len(lines)} tracks for semantic text-to-audio
alignment...")
print("-" * 60)

for line in tqdm(lines, desc="Computing CLAP Vectors"):
```

```
entry = json.loads(line)

track_id = entry["id"]
audio_path = entry["path"]
text_prompt = entry["description"]

try:
    # Loading the waveform and its native sampling rate
    wav, sr = torchaudio.load(audio_path)

    # If the audio is stereo or multi-channel, average across
    channels to get a single channel
    if wav.shape[0] > 1:
        wav = wav.mean(dim=0, keepdim=True)

    # LAION-CLAP requires a strict 48kHz sample rate; resample if
    the audio deviates
    if sr != 48000:
        resampler = torchaudio.transforms.Resample(orig_freq=sr,
new_freq=48000)
        wav = resampler(wav)

    # Converting back to a 1D NumPy array as expected by the
    Hugging Face processor
    audio_numpy = wav.squeeze(0).numpy()

    # Tokenizing the text prompt and extracting audio spectrogram
    features simultaneously
    inputs = processor(
        text=[text_prompt],
        audio=audio_numpy,
        sampling_rate=48000,
        return_tensors="pt",
        padding=True
    )

    # Moving all tensors (input_ids, attention_mask,
    input_features) to the active device (GPU/CPU)
    inputs = {k: v.to(device) for k, v in inputs.items()}

    # Disabling gradient computation to optimize memory usage and
    speed up inference
```

```
with torch.no_grad():
    # Passing text tokens through the text encoder
    text_outputs = clap_model.get_text_features(
        input_ids=inputs["input_ids"],
        attention_mask=inputs["attention_mask"]
    )
    # Passing audio features through the audio encoder
    audio_outputs = clap_model.get_audio_features(
        input_features=inputs["input_features"],
        is_longer=inputs.get("is_longer", None)
    )

    # Handling structural variations in output format to extract
    raw embedding vectors
    text_embeds = text_outputs.pooler_output if
hasattr(text_outputs, 'pooler_output') else text_outputs
    audio_embeds = audio_outputs.pooler_output if
hasattr(audio_outputs, 'pooler_output') else audio_outputs

    # Projecting vectors onto the unit hypersphere. This ensures
    that the subsequent dot product calculates the exact Cosine Similarity
    between the text and audio spaces.
    text_embeds = F.normalize(text_embeds, p=2, dim=-1)
    audio_embeds = F.normalize(audio_embeds, p=2, dim=-1)

    # Computing the dot product of the normalized text and audio
    embeddings
    similarity = torch.matmul(text_embeds, audio_embeds.T).item()
    scores.append(similarity)

except Exception as e:
    print(f" Failed to process    {track_id}: {str(e)}")

# =====
# FINAL STATISTICAL BREAKDOWN
# =====
if scores:
    avg_clap = sum(scores) / len(scores)
    max_clap = max(scores)
    min_clap = min(scores)

    print("-" * 60)
```

```
print("? EMPIRICAL CLAP DATA FOR BATCH 1:")
print(f"✓ Average CLAP Cosine Similarity: {avg_clap:.4f}")
print(f"? Maximum Single Track Alignment: {max_clap:.4f}")
print(f" Minimum Single Track Alignment: {min_clap:.4f}")
print("-" * 60)

# Saving as clap_results1 to keep it distinct:
results_path = "/content/drive/MyDrive/audiocraft-
project/evaluation_batch1/clap_results1.json"
summary_data = {"average_clap": avg_clap, "max_clap": max_clap,
"min_clap": min_clap, "all_scores": scores}

with open(results_path, "w", encoding="utf-8") as rf:
    json.dump(summary_data, rf, indent=4)
print(f"? Metrics securely written to Google Drive
at:\n{results_path}")
```

Block 5: (FAD για Transformer)

```
import os
import glob
import numpy as np
import torch
import torchaudio
from transformers import AutoProcessor, ASTModel
from scipy.linalg import sqrtm
from tqdm import tqdm

torch.cuda.empty_cache()

device = "cuda" if torch.cuda.is_available() else "cpu"
print(f"? FAD Evaluation Engine Initialized on: {device.upper()}")

print("? Fetching pre-trained Audio Spectrogram Transformer
weights...")
model_name = "MIT/ast-finetuned-audioset-10-10-0.4593"
processor = AutoProcessor.from_pretrained(model_name)
ast_model = ASTModel.from_pretrained(model_name).to(device)
ast_model.eval()
```

```
real_audio_dir = "/content/drive/MyDrive/audiocraft-  
project/gtzan/Data/genres_original/disco"  
gen_audio_dir = "/content/drive/MyDrive/audiocraft-  
project/evaluation_batch1"  
  
real_files = glob.glob(os.path.join(real_audio_dir, "*.wav"))  
gen_files = [f for f in glob.glob(os.path.join(gen_audio_dir, "*.wav"))  
if "eval_track" in os.path.basename(f)]  
  
print(f"? Dataset Layout Located:")  
print(f" - Real Reference Set: {len(real_files)} studio tracks  
found.")  
print(f" - Generated Eval Set: {len(gen_files)} fine-tuned tracks  
found.")  
  
if len(real_files) == 0 or len(gen_files) == 0:  
    raise ValueError("✗ Error: One of your audio directories is missing  
files. Re-check your paths.")  
  
def extract_ast_embeddings(file_list, description_label):  
    embeddings = []  
  
    for path in tqdm(file_list, desc=f"Profiling {description_label}"):   
        try:  
            wav, sr = torchaudio.load(path)  
  
            # Downmixing to mono-channel  
            if wav.shape[0] > 1:  
                wav = wav.mean(dim=0, keepdim=True)  
  
            # Resampling to model specifications  
            # The MIT Audio Spectrogram Transformer (AST) natively  
            expects a 16kHz sampling rate  
            if sr != 16000:  
                resampler =  
torchaudio.transforms.Resample(orig_freq=sr, new_freq=16000)  
                wav = resampler(wav)  
  
            audio_numpy = wav.squeeze(0).numpy()  
  
            #Temporal truncation
```

```
# AST models evaluate fixed-duration spectrogram patches
(typically 10 seconds).
# Enforcing a 10-second ceiling (160,000 samples) to ensure
input shape uniformity.
max_samples = 16000 * 10
if len(audio_numpy) > max_samples:
    audio_numpy = audio_numpy[:max_samples]

# Spectrogram generation & padding: Converting raw 1D audio
array into log-mel filterbank features via the AST processor
inputs = processor(audio_numpy, sampling_rate=16000,
return_tensors="pt")
inputs = {k: v.to(device) for k, v in inputs.items()}

# Feature extraction & mean pooling
with torch.no_grad():
    outputs = ast_model(**inputs)
    # The model outputs a sequence of hidden states (batch,
sequence_length, hidden_dim). Compressing the temporal dimension via
mean pooling to derive a single static feature vector.
    pooled_output =
outputs.last_hidden_state.mean(dim=1).squeeze(0).cpu().numpy()
    embeddings.append(pooled_output)

except Exception as e:
    continue

return np.array(embeddings)

print("\n--- Running Multivariate Embedding Extractions ---")
real_embeds = extract_ast_embeddings(real_files, "Real Disco
Distribution")
gen_embeds = extract_ast_embeddings(gen_files, "Fine-Tuned Gen
Distribution")

print("\n--- Calculating Statistical Covariances ---")
# Fitting a multivariate Gaussian distribution over the embedding
clouds by computing the empirical mean vectors (mu) and covariance
matrices (sigma) for both domains.
mu_real = np.mean(real_embeds, axis=0)
mu_gen = np.mean(gen_embeds, axis=0)
```

```
sigma_real = np.cov(real_embeds, rowvar=False)
sigma_gen = np.cov(gen_embeds, rowvar=False)

#Fréchet Distance math execution: Calculating the geometric distance
between the two distributions in the embedding space.
# 1. Computing the squared L2 norm of the difference between mean
vectors.
mean_diff_squared = np.sum((mu_real - mu_gen) ** 2)

# 2. Computing the matrix square root of the product of the covariance
matrices.
cov_product_sqrt, _ = sqrtm(sigma_real.dot(sigma_gen), disp=False)

# Cleaning up imaginary numerical noise introduced during the matrix
square root calculation
if np.iscomplexobj(cov_product_sqrt):
    cov_product_sqrt = cov_product_sqrt.real

# 3. Computing the linear trace component of the Wasserstein-2 metric
trace_term = np.trace(sigma_real + sigma_gen - 2 * cov_product_sqrt)
final_fad_score = mean_diff_squared + trace_term

print("-" * 60)
print("? FINAL TRANS-FAD DATA FOR BATCH 1:")
print(f"✓ Calculated Fréchet Audio Distance: {final_fad_score:.4f}")
print("-" * 60)

# Saving as fad_results1 to keep it distinct:
results_path = "/content/drive/MyDrive/audiocraft-
project/evaluation_batch1/fad_results1.json"
summary_data = {
    "fad_score": float(final_fad_score),
    "mean_difference_component": float(mean_diff_squared),
    "trace_covariance_component": float(trace_term)
}
with open(results_path, "w", encoding="utf-8") as rf:
    json.dump(summary_data, rf, indent=4)
print(f"? Metrics securely written to Google Drive
at:\n{results_path}")
```

Block 6: (CLAP για Diffusion)

```
import os
import json
import torch
import torchaudio
import torch.nn.functional as F
from transformers import ClapModel, ClapProcessor
from tqdm import tqdm

# Clearing active hardware cache to prevent out-of-memory (OOM) errors
torch.cuda.empty_cache()

device = "cuda" if torch.cuda.is_available() else "cpu"
print(f"? Initializing CLAP Evaluation for Diffusion on:
{device.upper()}")

print(f"? Fetching pre-trained LAION-CLAP embedding models...")
model_name = "laion/clap-htsat-fused"
processor = ClapProcessor.from_pretrained(model_name)
clap_model = ClapModel.from_pretrained(model_name).to(device)
clap_model.eval() # Setting model to evaluation mode (disables dropout,
batchnorm updates)

manifest_path = "/content/drive/MyDrive/audiocraft-
project/evaluation_batch_diffusion/validation_manifest_diffusion.jsonl"

if not os.path.exists(manifest_path):
    raise FileNotFoundError(f"✗ Cannot find
validation_manifest_diffusion.jsonl at {manifest_path}")

scores = []
with open(manifest_path, "r", encoding="utf-8") as f:
    lines = [line.strip() for line in f if line.strip()]

print(f"? Processing {len(lines)} diffusion tracks for semantic
alignment...")
print("-" * 60)

for line in tqdm(lines, desc="Computing CLAP Vectors"):
    entry = json.loads(line)
```

```
track_id = entry["id"]
audio_path = entry["path"]
text_prompt = entry["description"]

try:
    # Loading the raw diffusion-generated waveform and its native
sampling rate
    wav, sr = torchaudio.load(audio_path)

    # Downmixing to mono-channel: If the generated track is stereo,
average across the channel dimension to enforce mono input
    if wav.shape[0] > 1:
        wav = wav.mean(dim=0, keepdim=True)

    # Resampling to model specifications: LAION-CLAP requires a
strict 48kHz sample rate; resample if the diffusion output deviates
    if sr != 48000:
        resampler = torchaudio.transforms.Resample(orig_freq=sr,
new_freq=48000)
        wav = resampler(wav)

    # Converting back to a 1D NumPy array as expected by the
Hugging Face processor pipeline
    audio_numpy = wav.squeeze(0).numpy()

    # Feature tokenization & extraction setup
    # Simultaneously tokenize the text prompt and extract audio
spectrogram features
    inputs = processor(
        text=[text_prompt],
        audio=audio_numpy,
        sampling_rate=48000,
        return_tensors="pt",
        padding=True
    )

    # Safely moving only valid tensor items to the active compute
hardware (GPU/CPU)
    inputs = {k: v.to(device) for k, v in inputs.items() if v is
not None}

    # Forward pass (joint embedding generation)
```

```
# Disabling gradient calculations to minimize VRAM usage and
optimize inference speed
with torch.no_grad():
    # VERSION-PROOF HUGGING FACE APPROACH:
    # Calling the model directly with **inputs. This
    automatically routes text tokens, audio features, and internal
    parameters like 'is_longer' to their respective encoders.
    outputs = clap_model(**inputs)

    # Extracting the raw, non-normalized embedding vectors from
    the joint text-audio space
    text_embeds = outputs.text_embeds
    audio_embeds = outputs.audio_embeds

    # L2 Normalization: Projecting vectors onto the unit
    hypersphere. This ensures that the subsequent dot product calculates
    the exact Cosine Similarity between the text and audio spaces.
    text_embeds = F.normalize(text_embeds, p=2, dim=-1)
    audio_embeds = F.normalize(audio_embeds, p=2, dim=-1)

    # Alignment calculation(cosine similarity): Computing the dot
    product of the normalized text and audio embeddings
    similarity = torch.matmul(text_embeds, audio_embeds.T).item()
    scores.append(similarity)

except Exception as e:
    print(f" Failed to process    {track_id}: {str(e)}")

# =====
# FINAL STATISTICAL BREAKDOWN
# =====
if scores:
    avg_clap = sum(scores) / len(scores)
    max_clap = max(scores)
    min_clap = min(scores)

    print("--" * 60)
    print(f"? EMPIRICAL CLAP DATA FOR DIFFUSION BATCH:")
    print(f"✔ Average CLAP Cosine Similarity: {avg_clap:.4f}")
    print(f"? Maximum Single Track Alignment: {max_clap:.4f}")
    print(f" Minimum Single Track Alignment: {min_clap:.4f}")
    print("--" * 60)
```

```
results_path = "/content/drive/MyDrive/audiocraft-  
project/evaluation_batch_diffusion/clap_results_diffusion.json"  
summary_data = {"average_clap": avg_clap, "max_clap": max_clap,  
"min_clap": min_clap, "all_scores": scores}  
  
with open(results_path, "w", encoding="utf-8") as rf:  
    json.dump(summary_data, rf, indent=4)  
    print(f"? Metrics written to Google Drive at:\n{results_path}")  
else:  
    print("✗ Error: No scores were collected. Check the individual  
track warnings above.")
```

Block 7: (FAD για Diffusion)

```
import os  
import glob  
import numpy as np  
import torch  
import torchaudio  
from transformers import AutoProcessor, ASTModel  
from scipy.linalg import sqrtm  
from tqdm import tqdm  
  
torch.cuda.empty_cache()  
  
device = "cuda" if torch.cuda.is_available() else "cpu"  
print(f"? FAD Evaluation for Diffusion Initialized on:  
{device.upper()}")  
  
print("? Fetching pre-trained Audio Spectrogram Transformer  
weights...")  
model_name = "MIT/ast-finetuned-audioset-10-10-0.4593"  
processor = AutoProcessor.from_pretrained(model_name)  
ast_model = ASTModel.from_pretrained(model_name).to(device)  
ast_model.eval()  
  
real_audio_dir = "/content/drive/MyDrive/audiocraft-  
project/gtzan/Data/genres_original/disco"
```

```
gen_audio_dir = "/content/drive/MyDrive/audiocraft-  
project/evaluation_batch_diffusion"  
  
real_files = glob.glob(os.path.join(real_audio_dir, "*.wav"))  
gen_files = [f for f in glob.glob(os.path.join(gen_audio_dir, "*.wav"))  
if "diff_track" in os.path.basename(f)]  
  
print(f"? Dataset Layout Located:")  
print(f" - Real Reference Set: {len(real_files)} tracks found.")  
print(f" - Diffusion Eval Set: {len(gen_files)} tracks found.")  
  
def extract_ast_embeddings(file_list, description_label):  
    embeddings = []  
  
    for path in tqdm(file_list, desc=f"Profiling {description_label}"):   
        try:  
            # Loading the raw audio waveform and extract its native  
            sampling rate  
            wav, sr = torchaudio.load(path)  
  
            # Downmixing to mono-channel: Average across channels if  
            the file is stereo to maintain a uniform 1D input shape  
            if wav.shape[0] > 1:  
                wav = wav.mean(dim=0, keepdim=True)  
  
            # Resampling to model specifications: The Audio Spectrogram  
            Transformer natively expects audio processed at 16kHz  
            if sr != 16000:  
                resampler =  
                torchaudio.transforms.Resample(orig_freq=sr, new_freq=16000)  
                wav = resampler(wav)  
  
            audio_numpy = wav.squeeze(0).numpy()  
  
            # Temporal truncation: AST evaluates fixed-size spectrogram  
            grids mapped to roughly 10 seconds of audio. Truncating any data  
            exceeding 10 seconds (160,000 samples) to preserve input consistency.  
            max_samples = 16000 * 10  
            if len(audio_numpy) > max_samples:  
                audio_numpy = audio_numpy[:max_samples]
```

```
# Spectrogram generation & padding : Converting raw time-
domain array into normalized log-mel filterbank features
inputs = processor(audio_numpy, sampling_rate=16000,
return_tensors="pt")
inputs = {k: v.to(device) for k, v in inputs.items()}

# Feature extraction & mean pooling: Disabling gradients
for speed and efficiency during forward pass evaluation
with torch.no_grad():
    outputs = ast_model(**inputs)
    # The raw output yields sequence-level token embeddings
(batch, sequence_length, hidden_dim). Executing mean pooling across the
sequence dimension (dim=1) to construct a static feature vector.
    pooled_output =
outputs.last_hidden_state.mean(dim=1).squeeze(0).cpu().numpy()
    embeddings.append(pooled_output)

except Exception as e:
    continue

return np.array(embeddings)

print("\n--- Running Multivariate Embedding Extractions ---")
real_embeds = extract_ast_embeddings(real_files, "Real Disco
Distribution")
gen_embeds = extract_ast_embeddings(gen_files, "Diffusion Gen
Distribution")

print("\n--- Calculating Statistical Covariances ---")
# Spatial distribution profiling: Model both high-dimensional embedding
clouds as multivariate Gaussian distributions by extracting their
empirical mean vectors (mu) and covariance matrices (sigma).
mu_real = np.mean(real_embeds, axis=0)
mu_gen = np.mean(gen_embeds, axis=0)

sigma_real = np.cov(real_embeds, rowvar=False)
sigma_gen = np.cov(gen_embeds, rowvar=False)

# Fréchet Distance math execution: Measuring the topological
discrepancy between the generated and real distribution spaces.
# 1. Computing the squared L2 norm distance between the spatial
centroids (means)
```

```
mean_diff_squared = np.sum((mu_real - mu_gen) ** 2)

# 2. Computing the matrix square root of the cross-covariance product
matrix
cov_product_sqrt, _ = sqrtm(sigma_real.dot(sigma_gen), disp=False)

# Dropping minor imaginary residuals caused by numerical floating-point
inaccuracies during sqrtm
if np.iscomplexobj(cov_product_sqrt):
    cov_product_sqrt = cov_product_sqrt.real

# 3. Summing the trace component elements to capture differences in
distribution spread and variance
trace_term = np.trace(sigma_real + sigma_gen - 2 * cov_product_sqrt)
final_fad_score = mean_diff_squared + trace_term

print("-" * 60)
print("? FINAL TRANS-FAD DATA FOR DIFFUSION:")
print(f"✔ Calculated Diffusion Fréchet Audio Distance:
{final_fad_score:.4f}")
print("-" * 60)

results_path = "/content/drive/MyDrive/audiocraft-
project/evaluation_batch_diffusion/fad_results_diffusion.json"
summary_data = {
    "fad_score": float(final_fad_score),
    "mean_difference_component": float(mean_diff_squared),
    "trace_covariance_component": float(trace_term)
}
with open(results_path, "w", encoding="utf-8") as rf:
    json.dump(summary_data, rf, indent=4)
print(f"? Metrics written to Google Drive at:\n{results_path}")
```

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης.

Διπλωματική Εργασία

162