Εφαρμογές αλγορίθμων γενετικής τεχνητής νοημοσύνης για την παραγωγή ηχητικού περιεχομένου στα νέα μέσα και την επικοινωνία

Αλέξανδρος Εμβολιάδης[1*] ,Πάρις Ξυλογιάννης[1], Νικόλαος Βρύζας[1,], Λάζαρος Βρύσης[1] , Χαράλαμπος Δημούλας[1]
[1] Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
[*]alemvoliadis@gmail.com

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία στοχεύει στην ανάλυση μεθόδων για την παραγωγή πολυμεσικών δεδομένων με ιδιαίτερη έμφαση στην επεξεργασία και παραγωγή ηχητικών σημάτων. Πιο συγκεκριμένα, η εργασία παρουσιάζει σύγχρονες τεχνικές βασισμένες σε αρχιτεκτονικές Βαθέων Νευρωνικών Δικτύων (ΒΝΔ) που στοχεύουν στην παραγωγή δεδομένων ήχου σε διάφορους τομείς (παραγωγή μουσικής, ομιλίας, περιβαλλοντικών ήχων και άλλα).Η εργασία ξεκινά με μια σύντομη ιστορική αναδρομή σε τεχνικές προσομοιώσεις καθώς και αναφορά σε σύγχρονες προσεγγίσεις. Συνεχίζει με την παρουσίαση νευραλγικών αρχιτεκτονικών ΒΝΔ, ικανές να παράξουν πληθώρα τύπων δεδομένων. Τέλος, η εργασία αναλύει μερικά σενάρια εφαρμογής για την παραγωγή ηχητικού περιεχομένου και παρουσιάζει τα συμπεράσματα.

Generative Artificial Intelligence applications for audio content production in new media and communication

ABSTRACT

This work aims to analyze methods for generating multimedia data, with a particular emphasis on the processing and production of audio signals. Specifically, the paper presents modern techniques based on Deep Neural Network (DNN) architectures that focus on generating audio data across various domains (such as music production, speech, environmental sounds, and more). The work begins with a brief historical overview of simulation techniques and references to contemporary approaches. It continues by presenting neural DNN architectures capable of producing a wide variety of data types. Finally, the paper analyzes several application scenarios for audio content generation and presents the conclusions.

1. Εισαγωγή

Οι εξελίξεις στη γενετική τεχνητή νοημοσύνη (Generative AI) έχουν προσελκύσει έντονο ενδιαφέρον από ακαδημαϊκούς, επιχειρηματίες και το ευρύ κοινό, κυρίως λόγω της ανάπτυξης ισχυρών μοντέλων (Foundational Models) με πρακτική εφαρμογή και διάθεση σε ευρεία χρήση. Παρότι αρχικά προσανατολισμένα στη μετατροπή κειμένου σε κείμενο, οι πολυτροπικές εφαρμογές, συμπεριλαμβανομένης της παραγωγής ηχητικού περιεχομένου, αναπτύσσονται ραγδαία.

1.1 Τεχνικές Προσομοίωσης και τεχνητή νοημοσύνη

Από τα μέσα του 20ού αιώνα, οι τεχνικές προσομοίωσης εξελίχθηκαν σημαντικά, με τη μέθοδο Monte Carlo [1] να επιλύει ντετερμινιστικά προβλήματα με πιθανοτικές διεργασίες. Παρά τις επιτυχίες της, η μέθοδος έχει περιορισμούς, όπως το υψηλό υπολογιστικό κόστος και την εξάρτηση από γεννήτριες τυχαίων αριθμών.

Οι σύγχρονες προσεγγίσεις ενσωματώνουν Βαθιά Νευρωνικά Δίκτυα (ΒΝΔ), τα οποία, μέσω πολλαπλών επιπέδων, αναλύουν και συνθέτουν δεδομένα υψηλής πολυπλοκότητας. Σημαντικές αρχιτεκτονικές είναι τα Variational Autoencoders (VAEs) [2], τα Generative Adversarial Networks (GANs) [3], και τα Diffusion Models [4], με καθένα να προσφέρει μοναδικά πλεονεκτήματα και προκλήσεις στην παραγωγή συνθετικών δεδομένων.

1.3 Αρχιτεκτονικές ΒΝΔ για γέννηση δεδομένων

Συγκεκριμένα, οι VAEs είναι αρχιτεκτονικές που βασίζονται στην κωδικοποίηση της εισόδου. Εισάγουν ένα πλήθος κατανομών λανθάνοντος χώρου και το μοντέλο εκπαιδεύεται στην ακριβή ανάκτηση της εισόδου, ενώ οι κατανομές προσεγγίζουν Γκαουσιανές κατανομές. Εφόσον εξαχθούν οι παράμετροι των κανονικών κατανομών είναι δυνατό να προσομοιωθεί ένα δείγμα του συνόλου εκπαίδευσης, λαμβάνοντας τυχαία δείγματα από τον λανθάνοντα χώρο. Τα GANs αποτελούνται από 2 αρχιτεκτονικές ΒΝΔ: την Γεννήτρια (Generator) και τον Επικριτή (Critic/Discriminator). Οι 2 αυτές αρχιτεκτονικές εκπαιδεύονται ταυτόχρονα, με την Γεννήτρια να λαμβάνει σαν είσοδο θόρυβο και να προσομοιώνει το σύνολο δεδομένων εκπαίδευσης, και τον Επικριτή να διακρίνει μεταξύ πραγματικών και συνθετικών δειγμάτων. Μέσα από αυτή τη διαδικασία, η Γεννήτρια είναι δυνατό να προσομοιώσει με μεγάλη ακρίβεια δείγματα του συνόλου εκπαίδευσης.

Ωστόσο, οι 2 παραπάνω αρχιτεκτονικές έχουν κι αυτές τους περιορισμούς τους. Οι VAEs είναι γνωστοί για την στάθμιση μεταξύ ποιότητας των συνθετικών δεδομένων και της ποικιλομορφίας τους [5], ενώ συχνά παράγουν θαμπά δείγματα. Όσων αφορά τα GANs, η Γεννήτρια είναι πιθανό να μάθει να παράγει συγκεκριμένα δείγματα από το σύνολο εκπαίδευσης (mode collapse) [5]. Τέλος, η εκπαίδευση των GANs χαρακτηρίζεται από μεγάλη αστάθεια λόγω της ανταγωνιστικής εκπαίδευσης [6].

Αυτούς τους περιορισμούς καλούνται να επιλύσουν τα μοντέλα Diffusion. Οι συγκεκριμένες αρχιτεκτονικές, λαμβάνουν σαν είσοδο δείγματα του συνόλου εκπαίδευσης και εισάγουν σταδιακά θόρυβο. Το τελικό δείγμα αποτελείται επί το

πλείστον από θόρυβο, και η αρχιτεκτονική εκπαιδεύεται στην πρόβλεψη του θορύβου σε κάθε στάδιο. Αυτή η διαδικασία αποθορυβοποίησης δίνει στα μοντέλα Diffusion ένα σημαντικό προβάδισμα στην σύνθεση ευκρινών και αναλυτικών δειγμάτων σε σύγκριση με τους VAEs και τα GANs. Ωστόσο, η συνεχής εισαγωγή θορύβου και σταδιακή αποθορυβοποίηση οδηγούν σε χρονοβόρα εκπαίδευση των συγκεκριμένων μοντέλων [5].

2. Εφαρμογές στην παραγωγή ηχητικού περιεχομένου

Η παρούσα ενότητα αναλύει σχολαστικά μερικές εφαρμογές της γενετικής τεχνητής νοημοσύνης για την παραγωγή ηχητικών δεδομένων. Αρκετές από τις περιπτώσεις χρησιμοποιούν πολυτροπικά δεδομένα σαν είσοδο για την σύνθεση ηχητικού περιεχομένου, κάτι που θα αναδειχθεί εντός της ενότητας.

2.1 Παραγωγή ομιλίας

Ένα από τα πρώτα προβλήματα που καταπιάστηκε η επιστημονική κοινότητα όσων αφορά την εφαρμογή της γενετικής τεχνητής νοημοσύνης, με πληθώρα υλοποιήσεων αλλά και εμπορικών προϊόντων, είναι η παραγωγή ομιλίας. Προσεγγίσεις τέτοιου τύπου δέχονται σαν είσοδο κείμενο ή/και φωνή για την παραγωγή λόγου βασισμένου στη φωνή και το κείμενο. Συνήθως, αυτές αποτελούνται από διαφορετικά στάδια. Αρχικά, το κείμενο και η φωνή τμηματοποιούνται και κάθε τμήμα επεξεργάζεται ξεχωριστά. Έπειτα, το σύστημα μετατρέπει το κείμενο σε φωνητικά στοιχεία, ενώ υπόκεινται σε φασματική ανάλυση. Τέλος, ο Vocoder είναι υπεύθυνος για την σύνθεση των τελικών ακουστικών κυμάτων. Σύγχρονες προσεγγίσεις υιοθετούν τη συγκεκριμένη διοχέτευση και εκμεταλλεύονται αρχιτεκτονικές που περιγράφονται στην ενότητα. Για παράδειγμα, η [7] εξελίσει τους VAEs διακριτοποιώντας τον λανθάνων χώρο (Vector Quantization). Εφαρμόζουν τη μέθοδο τους σε εικόνες και ομιλία. Συγκεκριμένα για την ομιλία, παρουσιάζουν πως το σύστημα μπορεί να παράξει τυχαία ομιλία και να πραγματοποιήσει μεταφορά ομιλητή, λαμβάνοντας σαν είσοδο αποκλειστικά ομιλία. Η [8] εκμεταλλεύεται ένα Diffusion μοντέλο για την παραγωγή κυματομορφής από κείμενο. Οι ενδιάμεσες φασματικές αναπαραστάσεις δίνονται σαν είσοδος στο Diffusion μοντέλο, που μαθαίνει, σύμφωνα με ένα δοσμένο κείμενο την αντήχηση των λέξεων.

Τέλος, άξια αναφοράς είναι η [9] όπου δέχεται σαν είσοδο κείμενο και δείγματα από κάποιον ομιλητή και παράγει ομιλία βάσει του δοσμένου κειμένου και ομιλητή. Χρησιμοποιώντας GANs, το κείμενο και τα δείγματα ομιλίας κωδικοποιούνται ξεχωριστά και συσχετίζονται για την παραγωγή ενός φασματογραφήματος, που μετατρέπεται σε κυματομορφή.

2.2 Παραγωγή Περιβαλλοντικών Ήχων

Η παραγωγή περιβαλλοντικών ήχων βρίσκει ευρεία πεδία εφαρμογής. Η σύνθεση δειγμάτων υψηλής ποιότητας, μπορεί να χρησιμοποιηθεί σε εφαρμογές επαυξημένης/εικονικής πραγματικότητας (AR/VR). Μέθοδοι που στοχεύουν στην γέννηση συνθετικών δειγμάτων βασίζονται σε γενικότερες μεθόδους γέννησης ήχου. Πάνω σε αυτήν την ιδέα βασίστηκαν οι [10]. Εκμεταλλεύονται έναν Autoencoder και ένα γλωσσικό μοντέλο (Language Model). Αφού εκπαιδευτεί το μοντέλο στο να

κάνει ακριβή ανάκτηση του αρχικού δείγματος, ήταν δυνατό να εκπαιδευτεί μια αυτοπαλίνδρομη διαδικασία ώστε να γεννά δείγματα του λανθάνοντος χώρου. Τέλος, εκπαιδεύοντας έναν κωδικοποιητή κειμένου, ήταν δυνατή η ανάπτυξη ενός μοντέλου που λαμβάνει κείμενο σαν είσοδο και παράγει ήχο. Κατά την ίδια κατεύθυνση κινείται και η [11], όπου η είσοδος στο σύστημα είναι ένα βίντεο. Το σύστημα αποτελείται από τρία στάδια. Αρχικά, κάθε καρέ του βίντεο κωδικοποιείται ξεχωριστά. Τα παραγόμενα διανύσματα συνδυάζονται σε ένα ίδιου μήκους με την ηχητική κυματομορφή. Το υπερ-διάνυσμα κωδικοποιείται περεταίρω και συνδυάζεται με την κωδικοποιημένη του έκδοση, όπου χρησιμοποιείται σαν είσοδος σε μία παραλλαγή ενός GAN.

Μια άλλη μέθοδος που χρησιμοποιείται για την σύνθεση δειγμάτων ηχητικών τοπίων λαμβάνει υπόψιν πολυτροπικά δεδομένα. Το σύστημα λαμβάνει σαν είσοδο βιντεο και ήχο. Κάθε τροπικότητα επεξεργάζεται ξεχωριστά. Εξάγονται λεζάντες για κάθε τροπικότητα ξεχωριστά. Οι λεζάντες χρησιμοποιούνται για να εκπαιδεύσουν ένα Μεγάλο Γλωσσικό Μοντέλο (Large Language Model) για να συσχετίσει και να συνδυάσει τις λεζάντες σε μία υπέρ-λεζάντα. Κάθε υπερ-λεζάντα χρησιμοποιείται σαν είσοδος σε ένα προεκπαιδευμένο σύστημα [12] για την συσχέτιση ήχου και υπερ-λεζάντας. Λαμβάνοντας το διάνυσμα από το κείμενο, ένα μοντέλο Diffusion είναι υπεύθυνο να προσομοιώνει το αντίστοιχο ηχητικό διάνυσμα. Αυτή χρησιμοποιείται για την προσομοίωση Φασματογραφημάτων ως είσοδοι σε ένα GAN που παράγει κυματομορφές.

2.3 Προσομοίωση Ακουστικής Χώρων

Η μέτρηση των ακουστικών χαρακτηριστικών ενός χώρου συχνά πραγματοποιείται με την καταγραφή της κρουστικής του απόκρισης (IR). Η προσομοίωση αυτών των χαρακτηριστικών αναβαθμίζει την εμπειρία του χρήστη σε εφαρμογές VR αλλά και μπορεί να χρησιμοποιηθεί για την ανάπτυξη ακουστικών φίλτρων.

Σύγχρονες μέθοδοι βασίζονται σε GANs και ανταγωνιστική εκπαίδευση. Για παράδειγμα η [13] χρησιμοποιεί ακουστικές παραμέτρους του χώρου, όπως ο χρόνος αντήχησης (Reverberation time, T_{60}), ο λόγος απευθείας και αντηχητικού σήματος (Direct-to-Reverberant Ratio), ο πρώιμος χρόνος εξασθένησης (Early Decay Time) και ο λόγος της πρώιμης προς την καθυστερημένη ηχητική ενέργεια (Early-to-Late Index) που εξάγονται απευθείας από τις κρουστικές αποκρίσεις των χώρων (RIRs). Αυτές οι παράμετροι χρησιμοποιούνται ως συνθήκη για την προσομοίωση κρουστικών αποκρίσεων.

Με την ανάδυση των πολυτροπικών μοντέλων εμφανίζονται και αντίστοιχες μέθοδοι για την προσομοίωση κρουστικών αποκρίσεων χώρων. Για παράδειγμα, η [14] λαμβάνει σαν είσοδο τις θέσεις ομιλητή και ακροατή, καθώς και την τρισδιάστατη αναπαράσταση του δωματίου σε μορφή πλέγματος (mesh). Αφού το πλέγμα απλοποιηθεί και κωδικοποιηθεί, συνδυάζεται με τις θέσεις ομιλητή/ακροατή και δίνεται σαν είσοδος σε μια γεννήτρια ενός GAN. Επιπλέον εισάγουν ως συνάρτηση αντικειμενικής απώλειας τη διαφορά μεταξύ του Διαγράμματος Αποσύνθεσης Ενέργειας (Energy Decay Relief) πραγματικής και συνθετικής κρουστικής απόκρισης.

Τέλος, ενδιαφέρον παρουσιάζει η [15]. Η συγκεκριμένη μέθοδος λαμβάνει σαν είσοδο μια φωτογραφία του δωματίου. Εξάγει το βάθος της εικόνας χρησιμοποιώντας ένα προ-εκπαιδευμένο μοντέλο [16]. Στη συνέχεια, η φωτογραφία

με το εκτιμώμενο βάθος συνδυάζονται και λειτουργούν ως είσοδος σε ένα προ-εκπαιδευμένο μοντέλο υπολογιστικής όρασης. Το διάνυσμα που προκύπτει, συνδυάζεται με θόρυβο και εισάγεται σε ένα GAN. Η μέθοδος παράγει κρουστικές αποκρίσεις υψηλής ποιότητας, ενώ είναι απλή στη χρήση από το ευρύ κοινό.

2.4 Παραγωγή Μουσικής

Η παραγωγή μουσικής είναι ένας ταχέως εξελισσόμενος τομέας της τεχνητής νοημοσύνης που αποσκοπεί στη δημιουργία, μετατροπή ή επεξεργασία της μουσικής με τη χρήση υπολογιστικών μοντέλων. Περιλαμβάνει μια ποικιλία εργασιών, καθεμία από τις οποίες ανταποκρίνεται σε διαφορετικές δημιουργικές ανάγκες.

Μια από τις πρώτες μεθόδους που ήταν επιτυχής στην παραγωγή μουσικής υψηλής ποιότητας, με ρυθμό δειγματοληψίας 48 kHz και χαμηλή καθυστέρηση, παρουσιάζεται στο [17]. Το συγκεκριμένο μοντέλο λειτουργεί με μία προσέγγιση δύο σταδίων: αρχικά ένα μοντέλο VAE εκπαιδεύεται ώστε να παράγει ποιοτικές αναπαραστάσεις των ηχητικών εισόδων, ενώ στην συνέχεια αξιοποιώντας ένα μοντέλο Επικριτή εκπαιδεύεται με ανταγωνιστική βελτιστοποίηση (adversarial fine-tuning) για τη περαιτέρω βελτίωση της ποιότητας του παραγόμενου ήχου. Κατ' αυτόν τον τρόπο καταφέρνει να μεταφέρει τον ήχο από ένα όργανο σε ένα άλλο ή να αλλάξει το ύφος της μουσικής.

Η μέθοδος [18] αποτελεί μία από τις πιο σύγχρονες προσεγγίσεις για την παραγωγή συνθετικής μουσικής. Σε αντίθεση με προηγούμενες μεθόδους πολλαπλών σταδίων, χρησιμοποιεί ένα γλωσσικό μοντέλο ενός μοναδικού σταδίου για την παραγωγή μουσικής υψηλής ποιότητας από κείμενα και μελωδίες. Το μοντέλο βασίζεται στον Υπολειμματικό Διανυσματικό Κβαντισμό (Residual Vector Quantization) της μεθόδου [19], επιτυγχάνοντας την παραγωγή μουσικών ηχητικών ροών σε υψηλή ανάλυση 32kHz.

Τέλος αξίζει να σημειωθεί ότι υπάρχουν αρκετές εμπορικές λύσεις, με την Suno AI [20] να αποτελεί ένα εξαιρετικά ενδιαφέρον παράδειγμα. Η εφαρμογή επιτρέπει στους χρήστες να δημιουργούν ολοκληρωμένα μουσικά κομμάτια, παρέχοντας μια συνεκτική περιγραφή για το παραγόμενο τραγούδι.

3. Εφαρμογές στα νέα μέσα και την επικοινωνία

Η διερεύνηση που παρουσιάστηκε σε σχέση με τις τελευταίες δυνατότητες στην παραγωγή ηχητικού περιεχομένου με χρήση μοντέλο γενετικής TN, μπορεί να αξιοποιηθεί στη γενικότερη προσπάθεια επέκτασης της εργαλειοθήκης των δημοσιογράφων και των παραγωγών περιεχομένου.

Η παραγωγή ομιλίας μπορεί να αξιοποιηθεί για την αποτελεσματικότερη δημιουργία εκφωνήσεων (voiceover) σε οπτικοακουστικό περιεχόμενο. Αυτό δεν είναι μόνο μία διέξοδος για την ενίσχυση των παραδοσιακών γραμμών εργασιών, αλλά μπορεί να διευκολύνει το άνοιγμα σε νέες μορφές περιεχομένου που είναι πιο προσιτές στις νεότερες γενιές και υποστηρίζονται από σύγχρονα μέσα κοινωνικής δικτύωσης. Παράλληλα μπορεί να αξιοποιηθεί σε συνδυασμό με συστήματα αυτόματης μετάφρασης για να διασφαλίσει την προσβασιμότητα της κοινότητας σε μορφές περιεχομένου όπου ως τώρα υπήρχαν γλωσσικοί περιορισμοί. Διασφαλίζεται επίσης έτσι η προσβασιμότητα σε πολυγλωσσικό περιεχόμενο σε άτομα με προβλήματα όρασης. Τα εργαλεία επεξεργασίας ήχου βάσει κειμένου (text-based

audio editing) αποτελούν μία πολύ ενδιαφέρουσα πτυχή που μπορεί να αλλάξει εντυπωσιακά τις υπάρχουσες ροές εργασιών. Η παραγωγή μουσικής μπορεί να λειτουργήσει επίσης ως καταλύτης σε νέες δημιουργικές πρακτικές για την υποστήριξη του παραγόμενου περιεχομένου, και τη δημιουργία εκπαιδευτικού και άλλου υλικού. Η παραγωγή περιβαλλοντικών ήχων και η προσομοίωση χώρων αποτελούν τεχνικές που αναμένεται να έχουν μεγάλη εφαρμογή στις διαδικασίες ηχητικού σχεδιασμού, αλλά και σε νέες μορφές περιεχομένου όπως η δημοσιογραφία εμβύθισης και οι εφαρμογές εκτεταμένης πραγματικότητας (xR) και εικονικής συν-τοποθέτησης (co-location), όπου οι συνδιαλεγόμενοι τοποθετούνται σε κοινό εικονικό χώρο.

Στην παρούσα φάση του πεδίου, είναι πολύ σημαντική η αξιολόγηση της ποιότητας των αποτελεσμάτων των παραπάνω τεχνικών μέσω πειραμάτων ακρόασης. Παράλληλα με τις νέες δυνατότητες, εγείρονται προβληματισμοί σχετικά με ηθικές και νομικές πτυχές του ζητήματος. Σχετικά ανοιχτά ζητήματα αφορούν την πνευματική ιδιοκτησία του παραγόμενου περιεχομένου. Επιπλέον, η ακεραιότητα του περιεχομένου που έχει παραχθεί με μοντέλα ΓΤΝ και η ενημέρωση του κοινού για τον τρόπο παραγωγής του περιεχομένου. Τέλος, σημαντικά είναι τα εργασιακά ζητήματα, που αφορούν την κατάργηση θέσεων εργασίας, αλλά και τους προβληματισμούς και δυσπιστία των εργαζομένων σε σχέση με την απόκτηση δεξιοτήτων και ψηφιακού γραμματισμού.

4. Συμπεράσματα

Η παρούσα εργασία έχει ως στόχο να παρουσιάσει εφαρμογές της γενετικής τεχνητής νοημοσύνης στην παραγωγή ηχητικών δεδομένων. Παρουσιάστηκαν διάφορες αρχιτεκτονικές ΒΝΔ με τα προτερήματα και περιορισμούς τους αλλά και ο τρόπος με τον οποίο ενσωματώνονται σε συστήματα παραγωγής ομιλίας, περιβαλλοντικών ήχων, κρουστικές αποκρίσεις χώρων και μουσικής. Από την παραπάνω ανάλυση γίνεται φανερό πως το πεδίο της γενετικής τεχνητής νοημοσύνης είναι ραγδαία εξελισσόμενο και παρουσιάζει πλήθος εφαρμογών. Η ανάδυση των ισχυρών μοντέλων δίνει τη δυνατότητα στην ανάπτυξη μεθοδολογιών χωρίς την απαίτηση εκπαίδευσης από την αρχή. Οι εξελίξεις στον κλάδο της επεξεργασίας πολυτροπικών δεδομένων επιδρούν θετικά στην γενετική τεχνητή νοημοσύνη, επιτρέποντας την ανάπτυξη εύρωστων μοντέλων. Τέλος, γίνεται φανερό πως η συνεισφορά όλων αυτών, οδηγεί σε σύνθετες αρχιτεκτονικές και μεθοδολογίες, που όμως είναι ικανές για την επίλυση πολλαπλών προβλημάτων ταυτόχρονα.

8. Αναφορές

[1] Benov, D. M. (2016). The Manhattan Project, the first electronic computer and the Monte Carlo method. *Monte Carlo Methods and Applications*, *22*(1), 73-79.

[2] Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

[4] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840-6851.

[5] Vivekananthan, S. (2024). Comparative Analysis of Generative Models: Enhancing Image Synthesis with VAEs, GANs, and Stable Diffusion. *arXiv preprint arXiv:2408.08751*.

[6] Ahmad, Z., Jaffri, Z. U. A., Chen, M., & Bao, S. (2024). Understanding GANs: fundamentals, variants, training challenges, applications, and open problems. *Multimedia Tools and Applications*, 1-77.

[7] Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, *30*.

[8] Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., & Kudinov, M. (2021, July). Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning* (pp. 8599-8608). PMLR.

[9] Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., & Ponti, M. A. (2022, June). Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning* (pp. 2709-2720). PMLR.

[10] Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., ... & Adi, Y. (2022). Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

[11]Liu, S., Li, S., & Cheng, H. (2021). Towards an end-to-end visual-to-raw-audio generation with GAN. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(3), 1299-1312.

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

[13] Ratnarajah, A., Tang, Z., & Manocha, D. (2020). IR-GAN: Room impulse response generator for far-field speech recognition. *arXiv preprint arXiv:2010.13219*.

[14] Ratnarajah, A., Tang, Z., Aralikatti, R., & Manocha, D. (2022, October). Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 924-933).

[15] Singh, N., Mentch, J., Ng, J., Beveridge, M., & Drori, I. (2021). Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 286-295).

[16] Tosi, F., Aleotti, F., Poggi, M., & Mattoccia, S. (2019). Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9799-9809).

[17] Caillon, A., & Esling, P. (2021). RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*.

[18] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., ... & Défossez, A. (2024). Simple and controllable music generation. *Advances in Neural Information Processing Systems*, *36*.

[19] Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. (2022). High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

[20] https://suno.com/