

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

Μεταπτυχιακή Εξειδίκευση στα Πληροφοριακά Συστήματα (MSc)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σύνθεση Φωνής με Χρήση Προεκπαιδευμένων
Μοντέλων της Πλατφόρμας Hugging Face

Αναστασιάδης Εμμανουήλ

Επιβλέπων Καθηγητής-Σύμβουλος:

Κωτσιαντής Σωτήριος



Πάτρα, Σεπτέμβριος 2026

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει τη σύνθεση φωνής (Text-to-Speech – TTS) μέσω της αξιοποίησης προεκπαιδευμένων νευρωνικών μοντέλων, με στόχο τη συγκριτική αξιολόγηση της ακουστικής ποιότητας και της υπολογιστικής αποδοτικότητας. Στο πλαίσιο της μελέτης αναπτύχθηκε ένα πειραματικό περιβάλλον βασισμένο σε εργαλεία ανοιχτού λογισμικού, το οποίο επιτρέπει την παραγωγή και ανάλυση συνθετικής ομιλίας υπό ελεγχόμενες συνθήκες. Η πειραματική διαδικασία επικεντρώθηκε σε δύο μοντέλα, το MMS_TTS_VITS και το SpeechT5, τα οποία αξιολογήθηκαν χρησιμοποιώντας κοινό σύνολο εισόδων και ενιαίο pipeline μετασχηματισμού από κείμενο σε ηχητική κυματομορφή. Η αξιολόγηση βασίστηκε σε συνδυασμό αντικειμενικών και proxy μετρικών, όπως το Mean Opinion Score (MOS), ο δείκτης Real-Time Factor (RTF), καθώς και φασματικά χαρακτηριστικά όπως RMS energy, spectral centroid και MFCC.

Τα αποτελέσματα καταδεικνύουν ότι το MMS_TTS_VITS υπερτερεί σημαντικά ως προς τη φυσικότητα και τη συνοχή της παραγόμενης ομιλίας, επιτυγχάνοντας υψηλότερες τιμές MOS και πιο ισορροπημένη φασματική κατανομή. Αντιθέτως, το SpeechT5 εμφανίζει καλύτερη υπολογιστική αποδοτικότητα, αλλά μειωμένη ακουστική ποιότητα. Η σύγκριση αυτή αναδεικνύει το βασικό trade-off μεταξύ ποιότητας και ταχύτητας που χαρακτηρίζει τα σύγχρονα συστήματα TTS. Τέλος, η εργασία υπογραμμίζει τη σημασία των προεκπαιδευμένων μοντέλων ως εργαλείων ταχείας ανάπτυξης, ενώ προτείνονται μελλοντικές κατευθύνσεις όπως η πολυγλωσσική σύνθεση φωνής και η real-time υλοποίηση.

Λέξεις-κλειδιά: σύνθεση φωνής, Text-to-Speech, προεκπαιδευμένα μοντέλα, νευρωνικά δίκτυα, MOS, RTF, φασματική ανάλυση, MMS_TTS_VITS, SpeechT5

Abstract

This thesis investigates Text-to-Speech (TTS) systems using pretrained neural models, focusing on the comparative evaluation of acoustic quality and computational

efficiency. An experimental framework was developed using open-source tools to generate and analyze synthetic speech under controlled conditions. The study evaluates two pretrained models, MMS_TTS_VITS and SpeechT5, using a unified input dataset and a consistent processing pipeline from text to audio waveform. The evaluation methodology combines objective and proxy metrics, including Mean Opinion Score (MOS), Real-Time Factor (RTF), and spectral features such as RMS energy, spectral centroid, and MFCC.

The results show that MMS_TTS_VITS significantly outperforms SpeechT5 in terms of naturalness and coherence, achieving higher MOS values and more balanced spectral characteristics. In contrast, SpeechT5 demonstrates better computational efficiency but lower audio quality. These findings highlight the trade-off between quality and performance in modern TTS systems. Finally, the study emphasizes the importance of pretrained models for rapid development and suggests future directions such as multilingual TTS, zero-shot voice cloning, and real-time inference optimization.

Keywords: Text-to-Speech, speech synthesis, pretrained models, neural networks, MOS, RTF, spectral analysis, MMS_TTS_VITS, SpeechT5

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή.....	5
1.1 Στόχοι και σκοπός της εργασίας.....	5
1.2 Κίνητρα και πεδίο εφαρμογών της σύνθεσης φωνής	7
1.3 Δομή της διπλωματικής εργασίας.....	10
Κεφάλαιο 2: Θεωρητικό Υπόβαθρο.....	11
2.1 Εξέλιξη της Σύνθεσης Φωνής (Text-to-Speech).....	11
2.2 Παραδοσιακές μέθοδοι (Concatenative, Parametric TTS)	14
2.3 Νευρωνικά δίκτυα και Deep Learning στην TTS	16
2.4 Εισαγωγή στα Προεκπαιδευμένα Μοντέλα	19
2.5 Η πλατφόρμα Hugging Face και το οικοσύστημά της.....	22
Κεφάλαιο 3: Τεχνολογίες και Εργαλεία.....	24
3.1 Προεκπαιδευμένα Μοντέλα για Σύνθεση Φωνής	24
3.2 Βιβλιοθήκες Hugging Face (Transformers, Datasets, Spaces)	28
3.3 Εργαλεία ανάπτυξης: Python, PyTorch/TensorFlow	32
3.4 Συστήματα αξιολόγησης φωνητικής σύνθεσης	37
Κεφάλαιο 4: Σχεδιασμός και Υλοποίηση	41
4.1 Αρχιτεκτονική του συστήματος.....	41
4.2 Επιλογή και περιγραφή προεκπαιδευμένων μοντέλων (π.χ. Tacotron2, VITS, SpeechT5).....	45
4.3 Διαδικασία fine-tuning και προσαρμογής μοντέλων	47
4.4 Ροή δεδομένων και pipeline (Text → Mel-spectrogram → Audio waveform).	48
4.5 Εργαλεία και περιβάλλον πειραμάτων	50
Κεφάλαιο 5: Πειραματική Αξιολόγηση.....	52
5.1 Περιγραφή συνόλων δεδομένων (π.χ. LJ Speech, Common Voice)	52
5.2 Μεθοδολογία αξιολόγησης (MOS, intelligibility, φυσικότητα)	55
5.3 Αποτελέσματα από τα προεκπαιδευμένα μοντέλα	56
5.4 Συγκριτική ανάλυση μεταξύ διαφορετικών μοντέλων.....	62
5.5 Συζήτηση και ερμηνεία αποτελεσμάτων.....	65
Κεφάλαιο 6: Συμπεράσματα και Μελλοντική Εργασία	67
6.1 Κύρια ευρήματα της εργασίας	67
6.2 Περιορισμοί και ανοιχτά ζητήματα.....	69

6.3 Μελλοντικές κατευθύνσεις (multilingual TTS, zero-shot voice cloning, real-time inference)	70
Βιβλιογραφία	71
Παράρτημα	76

Κεφάλαιο 1: Εισαγωγή

1.1 Στόχοι και σκοπός της εργασίας

Η ραγδαία πρόοδος των τεχνολογιών τεχνητής νοημοσύνης και ειδικότερα των νευρωνικών αρχιτεκτονικών βαθιάς μάθησης έχει αναδιαμορφώσει ριζικά το πεδίο της σύνθεσης φωνής, μετατρέποντάς το από ένα καθαρά μηχανιστικό πρόβλημα αναπαραγωγής ηχητικών κυματομορφών σε ένα σύνθετο πρόβλημα μάθησης αναπαραστάσεων υψηλής διάστασης που ενσωματώνουν σημασιολογικές, προσωδιακές και φωνητικές ιδιότητες. Στο πλαίσιο αυτό, η παρούσα εργασία στοχεύει στην εις βάθος διερεύνηση της αξιοποίησης προεκπαιδευμένων μοντέλων σύνθεσης φωνής μέσω της πλατφόρμας Hugging Face, με ιδιαίτερη έμφαση τόσο στις τεχνικές επιδόσεις όσο και στις υπολογιστικές και οικονομικές παραμέτρους που επηρεάζουν την υλοποίησή τους σε πραγματικά περιβάλλοντα. Ειδικότερα, η χρήση προεκπαιδευμένων μοντέλων μειώνει δραστικά το κόστος εκπαίδευσης, το οποίο σε πλήρως εποπτευόμενα σενάρια μπορεί να υπερβεί τις εκατοντάδες GPU-hours για datasets άνω των 20 ωρών ομιλίας, καθιστώντας την προσέγγιση transfer learning κρίσιμη για εφαρμογές χαμηλών πόρων (Liu, 2023).

Αρχικά, η εργασία επιδιώκει να προσδιορίσει με ακρίβεια τον ρόλο των σύγχρονων αρχιτεκτονικών TTS, όπως Tacotron2, VITS και diffusion-based μοντέλα, ως προς τη βελτιστοποίηση της ποιότητας του παραγόμενου λόγου και τη μείωση των σφαλμάτων ευθυγράμμισης μεταξύ κειμένου και φωνής. Πιο συγκεκριμένα, οι νεότερες end-to-end προσεγγίσεις παρουσιάζουν σημαντική μείωση στο Mel-Cepstral Distortion (MCD), το οποίο σε σύγχρονα συστήματα μπορεί να περιοριστεί κάτω από 4.0 dB, έναντι τιμών άνω των 5.5 dB σε παλαιότερες παραμετρικές μεθόδους (Yiwer et al., 2022). Επιπλέον, η ενσωμάτωση attention-based μηχανισμών και transformers επιτρέπει την

αποτελεσματικότερη μοντελοποίηση μακροπρόθεσμων εξαρτήσεων, μειώνοντας την πιθανότητα αποσυγχρονισμού (Latif et al., 2023).

Στη συνέχεια, κεντρικός στόχος αποτελεί η διερεύνηση της διαδικασίας μεταφοράς γνώσης (transfer learning) και προσαρμογής (fine-tuning) σε περιβάλλοντα περιορισμένων δεδομένων, όπου η διαθεσιμότητα επισημασμένων ηχητικών δεδομένων είναι περιορισμένη ή ανομοιογενής. Σε τέτοιες περιπτώσεις, τεχνικές few-shot learning έχουν επιδείξει αξιοσημείωτη απόδοση, επιτυγχάνοντας αποδεκτά επίπεδα φυσικότητας με λιγότερα από 30 λεπτά δεδομένων εκπαίδευσης, γεγονός που μειώνει δραστικά τόσο το κόστος συλλογής δεδομένων όσο και τον χρόνο ανάπτυξης (Liu, 2023). Παράλληλα, η αξιοποίηση συνθετικών δεδομένων για speaker adaptation έχει αποδειχθεί ότι μπορεί να επιταχύνει τη σύγκλιση των μοντέλων έως και 40%, χωρίς σημαντική υποβάθμιση της ποιότητας (Joshi & Garera, 2023).

Επιπροσθέτως, η εργασία στοχεύει στην ανάλυση της αρχιτεκτονικής pipeline σύνθεσης φωνής, η οποία συνήθως περιλαμβάνει τη μετατροπή κειμένου σε ενδιάμεση αναπαράσταση mel-spectrogram και στη συνέχεια σε ηχητική κυματομορφή. Η συγκεκριμένη διαδικασία αποτελεί κρίσιμο σημείο βελτιστοποίησης, καθώς η ποιότητα του mel-spectrogram επηρεάζει άμεσα την τελική ακουστική απόδοση. Σύγχρονες προσεγγίσεις που βασίζονται σε Continuous Wavelet Transform (CWT) έχουν επιδείξει βελτίωση στην αναπαράσταση των φασματικών χαρακτηριστικών, μειώνοντας τις απώλειες πληροφορίας σε υψηλές συχνότητες (Hu et al., 2024). Παράλληλα, τεχνικές όπως vector quantization και speech tokenization έχουν εισαχθεί για την αποτελεσματικότερη συμπίεση και αναπαράσταση της ομιλίας, επιτρέποντας πιο αποδοτική εκπαίδευση και inference (Chen et al., 2023; Bai et al., 2024).

Εξίσου σημαντικός στόχος της παρούσας μελέτης αφορά την αξιολόγηση της ποιότητας της συνθετικής φωνής μέσω αντικειμενικών και υποκειμενικών μετρικών. Η χρήση του Mean Opinion Score (MOS) παραμένει το κυρίαρχο κριτήριο αξιολόγησης, με σύγχρονα μοντέλα να επιτυγχάνουν τιμές άνω του 4.3 σε κλίμακα 5 βαθμών, προσεγγίζοντας την ανθρώπινη ομιλία (Maniati et al., 2022). Ωστόσο, η υποκειμενικότητα της MOS αξιολόγησης καθιστά αναγκαία τη συμπληρωματική χρήση μετρικών όπως PESQ και intelligibility scores, οι οποίες παρέχουν πιο αντικειμενική αποτίμηση της ακουστικής ποιότητας (Galdino et al., 2025). Επιπλέον, η ανάλυση προσωδίας και εκφραστικότητας αποκτά ιδιαίτερη σημασία, καθώς η απλή

ακουστική πιστότητα δεν επαρκεί για εφαρμογές υψηλής αλληλεπίδρασης (Barakat et al., 2024).

Η εργασία επιδιώκει να εξετάσει τις τεχνοοικονομικές διαστάσεις της ανάπτυξης συστημάτων TTS, οι οποίες συχνά παραβλέπονται στη βιβλιογραφία. Η μετάβαση από autoregressive σε non-autoregressive και diffusion-based μοντέλα έχει οδηγήσει σε σημαντική μείωση του latency, με ορισμένα συστήματα να επιτυγχάνουν real-time factors μικρότερους του 0.1, γεγονός που καθιστά εφικτή τη χρήση τους σε real-time εφαρμογές (Gao et al., 2024). Παράλληλα, η χρήση ενοποιημένων αρχιτεκτονικών όπως το VITS μειώνει τον αριθμό των απαιτούμενων υπομονάδων, περιορίζοντας την πολυπλοκότητα ανάπτυξης και το κόστος συντήρησης (Aghalarov et al., 2025). Επιπλέον, η δυνατότητα deployment σε edge συσκευές ανοίγει νέες προοπτικές για αποκεντρωμένες εφαρμογές, μειώνοντας το λειτουργικό κόστος και ενισχύοντας την ιδιωτικότητα των χρηστών (Mbonimpa et al., 2025).

Ακόμη, ένας σημαντικός άξονας της εργασίας αφορά τη διερεύνηση της εκφραστικότητας και της προσωποποίησης της συνθετικής φωνής, ιδιαίτερα μέσω τεχνικών όπως style transfer και voice cloning. Τα σύγχρονα diffusion-based μοντέλα έχουν επιτύχει επίπεδα εκφραστικότητας που προσεγγίζουν την ανθρώπινη ομιλία, επιτρέποντας τη μοντελοποίηση συναισθηματικών αποχρώσεων και προσωδιακών χαρακτηριστικών με υψηλή ακρίβεια (Kishore, 2025). Παράλληλα, προσεγγίσεις όπως StyleTTS και Llama-VITS ενσωματώνουν σημασιολογική πληροφορία, βελτιώνοντας τη συνοχή μεταξύ κειμένου και παραγόμενης ομιλίας (Han & Mesgarani, 2022; Feng & Yoshimoto, 2024).

1.2 Κίνητρα και πεδίο εφαρμογών της σύνθεσης φωνής

Η διαρκώς εντεινόμενη απαίτηση για φυσική, απρόσκοπτη και υψηλής πιστότητας αλληλεπίδραση μεταξύ ανθρώπου και υπολογιστικών συστημάτων έχει αναδείξει τη σύνθεση φωνής ως έναν από τους πλέον κρίσιμους πυλώνες της σύγχρονης τεχνητής νοημοσύνης. Υπό αυτό το πρίσμα, η ανάπτυξη προηγμένων συστημάτων Text-to-Speech (TTS) δεν αποτελεί απλώς ένα τεχνολογικό επίτευγμα, αλλά συνιστά βασική προϋπόθεση για την ενσωμάτωση ευφώνων υπηρεσιών σε καθημερινά ψηφιακά περιβάλλοντα. Η αυξανόμενη πολυπλοκότητα των εφαρμογών, σε συνδυασμό με τη

ζήτηση για προσωποποιημένη εμπειρία χρήστη, έχει οδηγήσει στην υιοθέτηση νευρωνικών αρχιτεκτονικών που υπερβαίνουν τις παραδοσιακές μεθόδους, επιτυγχάνοντας επίπεδα φυσικότητας που προσεγγίζουν την ανθρώπινη ομιλία (Tan et al., 2021).

Από διαφορετική σκοπιά, τα κίνητρα για την εξέλιξη της σύνθεσης φωνής ενισχύονται σημαντικά από τις ανάγκες προσβασιμότητας, όπου η τεχνολογία αυτή λειτουργεί ως καταλύτης κοινωνικής ενσωμάτωσης. Συγκεκριμένα, για άτομα με προβλήματα όρασης ή κινητικές δυσκολίες, τα συστήματα TTS παρέχουν κρίσιμες λειτουργίες, όπως η ανάγνωση ψηφιακού περιεχομένου και η αλληλεπίδραση με έξυπνες συσκευές, μειώνοντας το ψηφιακό χάσμα. Η ανάπτυξη τέτοιων εφαρμογών έχει καταστεί ιδιαίτερα αποδοτική χάρη στη χρήση προεκπαιδευμένων μοντέλων, τα οποία επιτρέπουν την προσαρμογή σε διαφορετικές γλώσσες και φωνητικά χαρακτηριστικά με περιορισμένα δεδομένα (Kumar et al., 2023). Επιπροσθέτως, η υποστήριξη γλωσσών χαμηλών πόρων μέσω crowdsourced δεδομένων και transfer learning τεχνικών ενισχύει την παγκόσμια διάδοση των TTS συστημάτων (Katumba et al., 2025).

Η εμπορική αξιοποίηση της σύνθεσης φωνής έχει γνωρίσει εκθετική ανάπτυξη, με εφαρμογές που εκτείνονται από εικονικούς βοηθούς και συστήματα εξυπηρέτησης πελατών έως πλατφόρμες παραγωγής πολυμέσων και gaming. Σε αυτά τα περιβάλλοντα, η ποιότητα της συνθετικής ομιλίας επηρεάζει άμεσα την εμπειρία του χρήστη και, κατ' επέκταση, την οικονομική απόδοση των συστημάτων. Ενδεικτικά, η βελτίωση της φυσικότητας της φωνής έχει συσχετιστεί με αύξηση της εμπλοκής χρηστών κατά ποσοστά που υπερβαίνουν το 20% σε εφαρμογές διαλογικών συστημάτων (Barakat et al., 2024). Παράλληλα, η μείωση του latency μέσω non-autoregressive μοντέλων έχει καταστήσει εφικτή την ανάπτυξη real-time υπηρεσιών, οι οποίες απαιτούν απόκριση σε κλίμακα χιλιοστών του δευτερολέπτου (Gao et al., 2024).

Εξίσου σημαντική διάσταση αποτελεί η αξιοποίηση της σύνθεσης φωνής σε περιβάλλοντα εκπαίδευσης και παραγωγής περιεχομένου, όπου η αυτοματοποίηση της αφήγησης επιτρέπει την ταχεία δημιουργία εκπαιδευτικού υλικού και πολυμεσικών εφαρμογών. Η χρήση τεχνικών expressive speech synthesis, που ενσωματώνουν συναισθηματικές και προσωδιακές παραμέτρους, οδηγεί σε πιο ελκυστικό και

αποτελεσματικό εκπαιδευτικό περιεχόμενο (Zhao & Yang, 2023). Παράλληλα, η ενσωμάτωση πολυτροπικών δεδομένων, όπως η ταυτόχρονη σύνθεση φωνής και χειρονομιών, ενισχύει τη ρεαλιστικότητα των εφαρμογών, ιδιαίτερα σε εικονικά περιβάλλοντα και συστήματα εικονικής πραγματικότητας (Wang et al., 2021).

Από μια πιο τεχνική οπτική, τα σύγχρονα συστήματα TTS αξιοποιούν πολύπλοκες αναπαραστάσεις της ομιλίας, όπως latent speech embeddings και mel-spectrograms, τα οποία επιτρέπουν την ακριβή μοντελοποίηση των φωνητικών χαρακτηριστικών. Η χρήση τεχνικών όπως speech tokenization και vector quantization έχει οδηγήσει σε σημαντική βελτίωση της αποδοτικότητας των μοντέλων, μειώνοντας τον όγκο των απαιτούμενων δεδομένων χωρίς απώλεια ποιότητας (Bai et al., 2024; Chen et al., 2023). Επιπλέον, η ανάπτυξη diffusion-based μοντέλων έχει επιτρέψει τη δημιουργία φωνών υψηλής πιστότητας, με μειωμένο θόρυβο και αυξημένη εκφραστικότητα (Kishore, 2025).

Η δυνατότητα προσωποποίησης της φωνής μέσω τεχνικών voice cloning και speaker adaptation έχει δημιουργήσει νέες επιχειρηματικές ευκαιρίες, αλλά και σημαντικές προκλήσεις. Από τη μία πλευρά, η ταχεία προσαρμογή σε νέα φωνητικά προφίλ με λίγα δεδομένα επιτρέπει την ανάπτυξη εξατομικευμένων εφαρμογών, όπως ψηφιακοί βοηθοί με μοναδική φωνητική ταυτότητα. Από την άλλη πλευρά, η κατάχρηση τέτοιων τεχνολογιών μπορεί να οδηγήσει σε ζητήματα ασφάλειας και προστασίας προσωπικών δεδομένων, καθιστώντας απαραίτητη την ανάπτυξη μηχανισμών ανίχνευσης και αποτροπής μη εξουσιοδοτημένης χρήσης (Zhang et al., 2023). Η εφαρμογή της σύνθεσης φωνής σε πολυγλωσσικά περιβάλλοντα και σε σενάρια code-mixed γλώσσας αναδεικνύει την ανάγκη για μοντέλα που μπορούν να χειριστούν πολύπλοκες γλωσσικές δομές. Η χρήση transformer-based αρχιτεκτονικών έχει αποδειχθεί ιδιαίτερα αποτελεσματική στην αντιμετώπιση αυτών των προκλήσεων, επιτρέποντας την ταυτόχρονη μοντελοποίηση διαφορετικών γλωσσικών επιπέδων (Latif et al., 2023). Παράλληλα, η ανάπτυξη ημι-εποπτευόμενων και αυτοεποπτευόμενων τεχνικών μάθησης μειώνει την εξάρτηση από επισημασμένα δεδομένα, επιταχύνοντας την ανάπτυξη νέων συστημάτων (Tu et al., 2020).

1.3 Δομή της διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία οργανώνεται σε μια συνεκτική αλληλουχία κεφαλαίων, τα οποία δομούνται με τρόπο που επιτρέπει τη σταδιακή εμβάθυνση από το θεωρητικό υπόβαθρο έως την πειραματική αξιολόγηση και τη σύνθεση συμπερασμάτων. Η συγκεκριμένη διάρθρωση δεν είναι τυχαία, αλλά αντανακλά τη μεθοδολογική προσέγγιση που απαιτείται για την ολοκληρωμένη μελέτη πολύπλοκων συστημάτων όπως τα μοντέλα σύνθεσης φωνής.

Αρχικά, το πρώτο κεφάλαιο λειτουργεί ως εισαγωγικό πλαίσιο, στο οποίο παρουσιάζονται οι στόχοι, τα κίνητρα και το γενικότερο πεδίο εφαρμογών της σύνθεσης φωνής, θέτοντας τις βάσεις για την κατανόηση των επόμενων εννοιών. Στη συνέχεια, το δεύτερο κεφάλαιο εστιάζει στο θεωρητικό υπόβαθρο, εξετάζοντας την εξέλιξη των TTS συστημάτων από τις παραδοσιακές concatenative και parametric μεθόδους έως τις σύγχρονες νευρωνικές προσεγγίσεις, οι οποίες βασίζονται σε βαθιά μάθηση και προεκπαιδευμένα μοντέλα (Mehri et al., 2023). Η ανάλυση αυτή περιλαμβάνει την παρουσίαση των βασικών αρχιτεκτονικών, καθώς και των τεχνικών που έχουν συμβάλει στη βελτίωση της ποιότητας της συνθετικής ομιλίας. Ακολούθως, το τρίτο κεφάλαιο επικεντρώνεται στις τεχνολογίες και τα εργαλεία που χρησιμοποιούνται για την ανάπτυξη συστημάτων TTS, με ιδιαίτερη έμφαση στις βιβλιοθήκες της πλατφόρμας Hugging Face και στα frameworks βαθιάς μάθησης. Η επιλογή των συγκεκριμένων εργαλείων βασίζεται τόσο στην ευελιξία όσο και στην υπολογιστική τους αποδοτικότητα, χαρακτηριστικά που είναι απαραίτητα για την υλοποίηση σύγχρονων εφαρμογών (Ma et al., 2025).

Στη συνέχεια, το τέταρτο κεφάλαιο παρουσιάζει τον σχεδιασμό και την υλοποίηση του προτεινόμενου συστήματος, περιγράφοντας αναλυτικά την αρχιτεκτονική, τη ροή δεδομένων και τη διαδικασία fine-tuning των μοντέλων. Η ενότητα αυτή περιλαμβάνει επίσης την ανάλυση του pipeline σύνθεσης φωνής, το οποίο αποτελείται από διαδοχικά στάδια επεξεργασίας που μετατρέπουν το κείμενο σε ηχητική κυματομορφή (Yiwere et al., 2022). Ακολούθως, το πέμπτο κεφάλαιο επικεντρώνεται στην πειραματική αξιολόγηση, όπου παρουσιάζονται τα χρησιμοποιούμενα σύνολα δεδομένων, η μεθοδολογία αξιολόγησης και τα αποτελέσματα των πειραμάτων. Η ανάλυση αυτή

περιλαμβάνει τόσο αντικειμενικές όσο και υποκειμενικές μετρικές, επιτρέποντας μια ολοκληρωμένη αποτίμηση της απόδοσης των μοντέλων (Maniati et al., 2022).

Κεφάλαιο 2: Θεωρητικό Υπόβαθρο

2.1 Εξέλιξη της Σύνθεσης Φωνής (Text-to-Speech)

Η ιστορική εξέλιξη της σύνθεσης φωνής συνιστά ένα πολυσύνθετο και πολυεπίπεδο ερευνητικό πεδίο, το οποίο αντανakλά τη διαχρονική προσπάθεια της επιστημονικής κοινότητας να γεφυρώσει το χάσμα μεταξύ συμβολικής γλωσσικής αναπαράστασης και ακουστικής παραγωγής λόγου. Αρχικά, οι πρώτες προσεγγίσεις σύνθεσης φωνής βασίζονταν σε αυστηρά καθορισμένα κανόνες και φωνητικά μοντέλα, τα οποία επιχειρούσαν να προσομοιώσουν τη λειτουργία του ανθρώπινου φωνητικού συστήματος μέσω παραμετρικών εξισώσεων. Ωστόσο, οι μέθοδοι αυτές παρουσίαζαν σημαντικούς περιορισμούς ως προς τη φυσικότητα και τη ρεαλιστικότητα της παραγόμενης ομιλίας, καθώς αδυνατούσαν να αποτυπώσουν τις πολύπλοκες προσωδιακές και συναισθηματικές μεταβολές που χαρακτηρίζουν τον ανθρώπινο λόγο (Tan et al., 2021).

Στη συνέχεια, η μετάβαση προς τις concatenative τεχνικές αποτέλεσε ένα κρίσιμο σημείο καμπής, καθώς η σύνθεση φωνής βασίστηκε στην ανασύνθεση πραγματικών ηχητικών αποσπασμάτων από εκτεταμένες βάσεις δεδομένων. Η προσέγγιση αυτή οδήγησε σε σημαντική βελτίωση της ακουστικής ποιότητας, δεδομένου ότι τα παραγόμενα σήματα προέρχονταν από φυσική ομιλία. Παρά ταύτα, η ανάγκη για μεγάλα και προσεκτικά επισημασμένα datasets, σε συνδυασμό με την περιορισμένη ευελιξία ως προς τη δημιουργία νέων φωνητικών μοτίβων, περιόρισε την επεκτασιμότητα των συστημάτων αυτών (Mehrish et al., 2023). Επιπλέον, η αδυναμία προσαρμογής σε νέα φωνητικά προφίλ χωρίς εκτενή επανεκπαίδευση καθιστούσε τις concatenative μεθόδους λιγότερο αποδοτικές σε δυναμικά περιβάλλοντα. Ακολούθως, η εισαγωγή των στατιστικών παραμετρικών μοντέλων σηματοδότησε μια σημαντική εξέλιξη, καθώς επέτρεψε την αναπαράσταση της ομιλίας μέσω πιθανοτικών κατανομών και παραμέτρων χαμηλής διάστασης. Τα μοντέλα αυτά, τα οποία βασίζονταν κυρίως σε Hidden Markov Models (HMM), προσέφεραν μεγαλύτερη

ευελιξία και δυνατότητα ελέγχου της παραγόμενης φωνής. Παρά τα πλεονεκτήματά τους, η ποιότητα της ομιλίας παρέμενε υποδεέστερη σε σύγκριση με τις concatenative προσεγγίσεις, κυρίως λόγω της υπεραπλοποίησης των φωνητικών χαρακτηριστικών και της απώλειας λεπτομερειών στο φασματικό πεδίο (Tan et al., 2021).

Η επόμενη καθοριστική φάση στην εξέλιξη της σύνθεσης φωνής συνδέεται άρρηκτα με την εμφάνιση των νευρωνικών δικτύων και, ειδικότερα, των αρχιτεκτονικών βαθιάς μάθησης. Η μετάβαση από τα παραδοσιακά στατιστικά μοντέλα σε νευρωνικά δίκτυα επέτρεψε την άμεση εκμάθηση μη γραμμικών σχέσεων μεταξύ κειμένου και ομιλίας, εξαλείφοντας την ανάγκη για χειροκίνητη εξαγωγή χαρακτηριστικών. Ειδικότερα, τα πρώτα νευρωνικά μοντέλα TTS, όπως το Tacotron, εισήγαγαν την έννοια της end-to-end εκπαίδευσης, όπου το σύστημα μαθαίνει απευθείας τη χαρτογράφηση από ακολουθίες χαρακτήρων σε φασματικές αναπαραστάσεις (Rai et al., 2024). Η προσέγγιση αυτή οδήγησε σε δραστική βελτίωση της φυσικότητας της ομιλίας, μειώνοντας παράλληλα την πολυπλοκότητα του pipeline.

Η ενσωμάτωση μηχανισμών attention αποτέλεσε καταλυτικό παράγοντα για την επιτυχία των νευρωνικών μοντέλων, καθώς επέτρεψε την αποτελεσματική ευθυγράμμιση μεταξύ εισόδου και εξόδου, αντιμετωπίζοντας προβλήματα όπως η παράλειψη ή επανάληψη φωνητικών μονάδων. Παράλληλα, η ανάπτυξη vocoders βασισμένων σε νευρωνικά δίκτυα, όπως το WaveGlow, συνέβαλε στη σημαντική βελτίωση της ποιότητας της τελικής ηχητικής εξόδου, επιτυγχάνοντας υψηλή πιστότητα και μειωμένο θόρυβο (Rai et al., 2024). Η συνδυασμένη χρήση acoustic models και neural vocoders κατέστησε δυνατή την παραγωγή ομιλίας με Mean Opinion Score που προσεγγίζει ή και υπερβαίνει το 4.0, πλησιάζοντας την ανθρώπινη απόδοση (Maniati et al., 2022). Στη σύγχρονη εποχή, η εξέλιξη των TTS συστημάτων χαρακτηρίζεται από τη μετάβαση σε πλήρως ενοποιημένες και μη αυτοπαλινδρομικές (non-autoregressive) αρχιτεκτονικές, οι οποίες προσφέρουν σημαντικά πλεονεκτήματα ως προς την ταχύτητα και την υπολογιστική αποδοτικότητα. Μοντέλα όπως το FastSpeech και το VITS επιτρέπουν την παράλληλη παραγωγή φωνητικών χαρακτηριστικών, μειώνοντας δραστικά τον χρόνο inference και καθιστώντας εφικτή τη χρήση τους σε εφαρμογές πραγματικού χρόνου (Aghalarov et al., 2025). Επιπλέον, τα μοντέλα αυτά ενσωματώνουν στοχαστικές διαδικασίες και variational μηχανισμούς, επιτρέποντας την παραγωγή πιο φυσικών και ποικίλων φωνητικών εκφράσεων.

Η εμφάνιση diffusion-based μοντέλων έχει εισαγάγει ένα νέο παράδειγμα στη σύνθεση φωνής, βασισμένο σε στοχαστικές διαδικασίες σταδιακής αποθορυβοποίησης. Τα μοντέλα αυτά έχουν επιδείξει εξαιρετική απόδοση ως προς την ποιότητα και την εκφραστικότητα της ομιλίας, επιτυγχάνοντας υψηλή πιστότητα ακόμη και σε συνθήκες περιορισμένων δεδομένων (Kishore, 2025). Επιπλέον, η δυνατότητα ελέγχου της διαδικασίας παραγωγής επιτρέπει τη λεπτομερή ρύθμιση χαρακτηριστικών όπως ο τόνος, η ένταση και ο ρυθμός ομιλίας.

Εξίσου σημαντική εξέλιξη αποτελεί η ενσωμάτωση transformer-based αρχιτεκτονικών, οι οποίες έχουν αποδειχθεί ιδιαίτερα αποτελεσματικές στην επεξεργασία ακολουθιών και στη μοντελοποίηση μακροπρόθεσμων εξαρτήσεων. Η χρήση transformers στη σύνθεση φωνής επιτρέπει την καλύτερη κατανόηση της σημασιολογίας του κειμένου, οδηγώντας σε πιο συνεκτική και φυσική ομιλία (Latif et al., 2023). Παράλληλα, η ανάπτυξη μοντέλων όπως το SpeechT5 και το Clip-TTS ενσωματώνει πολυτροπική πληροφορία, βελτιώνοντας τη συσχέτιση μεταξύ κειμένου και ηχητικής αναπαράστασης (Liu, 2025). Από μια διαφορετική οπτική, η αξιοποίηση προεκπαιδευμένων μοντέλων και τεχνικών μεταφοράς γνώσης έχει επιταχύνει σημαντικά την εξέλιξη των TTS συστημάτων, επιτρέποντας την ανάπτυξη εφαρμογών σε γλώσσες και περιβάλλοντα με περιορισμένα δεδομένα. Η δυνατότητα εκπαίδευσης μοντέλων με ελάχιστα δεδομένα, μέσω few-shot learning και semi-supervised τεχνικών, έχει ανοίξει νέους ορίζοντες για τη διάδοση της τεχνολογίας σε παγκόσμιο επίπεδο (Tu et al., 2020; Liu, R., et al., 2022). Παράλληλα, η χρήση synthetic data έχει συμβάλει στην αύξηση της ποικιλίας και της ποιότητας των datasets, βελτιώνοντας την απόδοση των μοντέλων (Joshi & Garera, 2023).

Η εξέλιξη της σύνθεσης φωνής δεν περιορίζεται μόνο στην τεχνολογική διάσταση, αλλά επεκτείνεται και σε ζητήματα ασφάλειας, ηθικής και ανίχνευσης συνθετικού περιεχομένου. Η ανάπτυξη datasets όπως το TIMIT-TTS και τεχνικών ανίχνευσης συνθετικής ομιλίας υπογραμμίζει την ανάγκη για υπεύθυνη χρήση της τεχνολογίας, ιδιαίτερα σε περιβάλλοντα όπου η αυθεντικότητα της φωνής αποτελεί κρίσιμο παράγοντα (Salvi et al., 2023). Επιπλέον, η πρόοδος στην αναγνώριση ταυτότητας φωνής μετά από ανασύνθεση αναδεικνύει τις προκλήσεις που σχετίζονται με την προστασία προσωπικών δεδομένων (Gaudier et al., 2024).

2.2 Παραδοσιακές μέθοδοι (Concatenative, Parametric TTS)

Η πρώτη φάση της σύνθεσης φωνής χαρακτηρίζεται από την ανάπτυξη μεθοδολογιών που επιδίωκαν την αναπαραγωγή της ανθρώπινης ομιλίας μέσω είτε άμεσης επαναχρησιμοποίησης ηχητικών τμημάτων είτε μέσω μαθηματικής μοντελοποίησης των φωνητικών χαρακτηριστικών. Στο πλαίσιο αυτό, οι παραδοσιακές μέθοδοι σύνθεσης φωνής, και ειδικότερα οι concatenative και parametric προσεγγίσεις, αποτέλεσαν τον θεμέλιο λίθο πάνω στον οποίο οικοδομήθηκαν τα σύγχρονα νευρωνικά συστήματα. Παρά τις σημαντικές τεχνολογικές εξελίξεις των τελευταίων ετών, η κατανόηση των συγκεκριμένων μεθόδων παραμένει κρίσιμη, καθώς αναδεικνύει τόσο τα βασικά προβλήματα όσο και τις σχεδιαστικές επιλογές που εξακολουθούν να επηρεάζουν την αρχιτεκτονική των σύγχρονων TTS συστημάτων (Tan et al., 2021). Αρχικά, η concatenative σύνθεση φωνής βασίζεται στην ιδέα της σύνθεσης νέας ομιλίας μέσω της συρραφής προϋπαρχόντων ηχητικών αποσπασμάτων, τα οποία έχουν καταγραφεί από φυσικούς ομιλητές και αποθηκευτεί σε μεγάλες βάσεις δεδομένων. Τα αποσπάσματα αυτά μπορεί να αντιστοιχούν σε φωνήματα, διφωνήματα ή ακόμα και ολόκληρες λέξεις, ανάλογα με τον σχεδιασμό του συστήματος. Η βασική αρχή λειτουργίας συνίσταται στην επιλογή της βέλτιστης ακολουθίας μονάδων που αντιστοιχούν στο προς σύνθεση κείμενο και στην ομαλή σύνδεσή τους, ώστε να ελαχιστοποιούνται οι ακουστικές ασυνέχειες. Ωστόσο, η διαδικασία αυτή συνεπάγεται υψηλή υπολογιστική πολυπλοκότητα κατά την αναζήτηση της βέλτιστης ακολουθίας, ιδιαίτερα όταν το σύστημα καλείται να επιλέξει μεταξύ εκατοντάδων χιλιάδων πιθανών συνδυασμών μονάδων.

Η ποιότητα της concatenative σύνθεσης εξαρτάται άμεσα από το μέγεθος και την ποικιλία του corpus εκπαίδευσης, γεγονός που δημιουργεί σημαντικές απαιτήσεις σε αποθηκευτικό χώρο και κόστος συλλογής δεδομένων. Ενδεικτικά, για την επίτευξη υψηλής φυσικότητας απαιτούνται συνήθως δεκάδες ώρες ηχογραφημένης ομιλίας, οι οποίες πρέπει να είναι επιμελώς επισημασμένες και ακουστικά συνεπείς. Παράλληλα, η αδυναμία γενίκευσης σε μη προβλεπόμενα φωνητικά μοτίβα περιορίζει τη δυνατότητα παραγωγής νέων εκφραστικών χαρακτηριστικών, καθιστώντας τα συστήματα αυτά λιγότερο ευέλικτα σε δυναμικά περιβάλλοντα (Mehrish et al., 2023). Συνεπώς, ενώ η concatenative προσέγγιση μπορεί να επιτύχει υψηλά επίπεδα

φυσικότητας υπό ιδανικές συνθήκες, η έλλειψη προσαρμοστικότητας αποτελεί σημαντικό μειονέκτημα.

Από μια διαφορετική οπτική, η παραμετρική σύνθεση φωνής επιχειρεί να αντιμετωπίσει τους περιορισμούς των concatenative μεθόδων μέσω της αναπαράστασης της ομιλίας ως σύνολο παραμέτρων που περιγράφουν τα φωνητικά χαρακτηριστικά του σήματος. Στα συστήματα αυτά, η παραγωγή ομιλίας βασίζεται σε στατιστικά μοντέλα, όπως τα Hidden Markov Models (HMM), τα οποία μαθαίνουν τις πιθανότητες μετάβασης μεταξύ φωνητικών καταστάσεων και τις αντίστοιχες κατανομές χαρακτηριστικών. Η προσέγγιση αυτή επιτρέπει την παραγωγή νέων φωνητικών συνδυασμών χωρίς την ανάγκη αποθήκευσης μεγάλων βάσεων δεδομένων, μειώνοντας σημαντικά το αποθηκευτικό και υπολογιστικό κόστος. Ωστόσο, η παραμετρική σύνθεση παρουσιάζει εγγενείς περιορισμούς ως προς την ποιότητα της παραγόμενης ομιλίας, καθώς η διαδικασία παραμετροποίησης οδηγεί σε απώλεια φασματικών λεπτομερειών και σε υπεραπλοποίηση των ακουστικών χαρακτηριστικών. Το φαινόμενο αυτό, γνωστό ως over-smoothing, έχει ως αποτέλεσμα την παραγωγή ομιλίας που συχνά χαρακτηρίζεται ως «μηχανική» ή «επίπεδη», στερούμενη της φυσικής μεταβλητότητας που χαρακτηρίζει την ανθρώπινη φωνή (Tan et al., 2021). Επιπλέον, η περιορισμένη ικανότητα μοντελοποίησης μη γραμμικών σχέσεων μεταξύ γλωσσικών και ακουστικών χαρακτηριστικών καθιστά δύσκολη την αναπαραγωγή σύνθετων προσωδιακών μοτίβων.

Οι δε παραμετρικές μέθοδοι προσφέρουν σημαντικά πλεονεκτήματα, ιδιαίτερα σε περιβάλλοντα με περιορισμένους πόρους. Η μικρότερη απαίτηση σε δεδομένα και η δυνατότητα ελέγχου της παραγόμενης ομιλίας μέσω παραμέτρων καθιστούν τα συστήματα αυτά κατάλληλα για εφαρμογές ενσωματωμένων συστημάτων και edge computing. Επιπλέον, η χαμηλή πολυπλοκότητα των μοντέλων επιτρέπει την υλοποίηση σε συσκευές με περιορισμένες υπολογιστικές δυνατότητες, γεγονός που ήταν ιδιαίτερα σημαντικό σε πρώιμες εφαρμογές TTS (Yiwere et al., 2022). Στη συνέχεια, η σύγκριση μεταξύ concatenative και parametric μεθόδων αναδεικνύει ένα σαφές trade-off μεταξύ ποιότητας και ευελιξίας. Ενώ οι concatenative προσεγγίσεις προσφέρουν υψηλότερη φυσικότητα λόγω της χρήσης πραγματικών ηχητικών δεδομένων, οι parametric μέθοδοι παρέχουν μεγαλύτερη δυνατότητα ελέγχου και προσαρμογής. Ωστόσο, καμία από τις δύο προσεγγίσεις δεν καταφέρνει να επιτύχει

ταυτόχρονα υψηλή ποιότητα, ευελιξία και αποδοτικότητα, γεγονός που οδήγησε στην ανάπτυξη υβριδικών μοντέλων και, τελικά, στη μετάβαση προς νευρωνικές αρχιτεκτονικές (Mehrish et al., 2023).

Επιπροσθέτως, η περιορισμένη ικανότητα των παραδοσιακών μεθόδων να υποστηρίξουν πολυγλωσσικά και πολυφωνικά σενάρια αποτελεί έναν ακόμη σημαντικό περιορισμό. Η προσαρμογή σε νέες γλώσσες ή ομιλητές απαιτεί συνήθως εκτενή επανεκπαίδευση ή συλλογή νέων δεδομένων, διαδικασία που συνεπάγεται σημαντικό κόστος και χρόνο ανάπτυξης. Αντίθετα, οι σύγχρονες προσεγγίσεις που βασίζονται σε μεταφορά γνώσης επιτρέπουν την ταχεία προσαρμογή με ελάχιστα δεδομένα, υπογραμμίζοντας τις αδυναμίες των παραδοσιακών μεθόδων (Liu, R., et al., 2022). Αξιοσημείωτο είναι επίσης το γεγονός ότι οι παραδοσιακές μέθοδοι παρουσιάζουν περιορισμούς στην αναπαραγωγή εκφραστικών χαρακτηριστικών, όπως το συναίσθημα και η προσωδία. Η έλλειψη δυναμικής μοντελοποίησης των συναισθηματικών παραμέτρων οδηγεί σε μονοτονική ομιλία, η οποία δεν είναι κατάλληλη για εφαρμογές που απαιτούν υψηλό βαθμό αλληλεπίδρασης, όπως τα διαλογικά συστήματα και οι εικονικοί βοηθοί (Barakat et al., 2024). Επιπλέον, η αδυναμία ενσωμάτωσης σημασιολογικής πληροφορίας περιορίζει την κατανόηση του περιεχομένου, οδηγώντας σε λιγότερο συνεκτική και φυσική ομιλία.

2.3 Νευρωνικά δίκτυα και Deep Learning στην TTS

Η εισαγωγή των νευρωνικών δικτύων και των τεχνικών βαθιάς μάθησης στο πεδίο της σύνθεσης φωνής σηματοδότησε μια θεμελιώδη μετατόπιση από τις παραδοσιακές προσεγγίσεις προς ένα πλήρως δεδομενοκεντρικό παράδειγμα, στο οποίο η γνώση δεν ενσωματώνεται μέσω χειροκίνητα σχεδιασμένων κανόνων, αλλά προκύπτει μέσω της εκμάθησης πολύπλοκων μη γραμμικών συναρτήσεων από μεγάλης κλίμακας δεδομένα. Η μετάβαση αυτή κατέστη εφικτή λόγω της εκθετικής αύξησης της υπολογιστικής ισχύος και της διαθεσιμότητας μεγάλων συνόλων δεδομένων, επιτρέποντας την εκπαίδευση βαθιών αρχιτεκτονικών με εκατομμύρια ή και δισεκατομμύρια παραμέτρους. Στο πλαίσιο αυτό, τα νευρωνικά μοντέλα TTS επιτυγχάνουν πλέον επίπεδα φυσικότητας που προσεγγίζουν την ανθρώπινη ομιλία, με τιμές Mean Opinion Score (MOS) που υπερβαίνουν το 4.3 σε κλίμακα πέντε βαθμών,

γεγονός που υποδηλώνει την υψηλή ποιότητα της παραγόμενης φωνής (Maniati et al., 2022). Αρχικά, η θεμελιώδης αρχή των νευρωνικών TTS συστημάτων βασίζεται στη μοντελοποίηση της συνάρτησης μετασχηματισμού από συμβολικές γλωσσικές αναπαραστάσεις σε ακουστικά χαρακτηριστικά, συνήθως μέσω της εκμάθησης μιας συνάρτησης $f: X \rightarrow Y$, όπου το X αντιστοιχεί σε ακολουθίες χαρακτήρων ή φωνημάτων και το Y σε φασματικές αναπαραστάσεις όπως τα mel-spectrograms. Η χρήση βαθιών νευρωνικών δικτύων επιτρέπει την προσέγγιση αυτής της συνάρτησης με υψηλή ακρίβεια, αξιοποιώντας πολλαπλά επίπεδα αφηρημένης αναπαράστασης. Ειδικότερα, τα convolutional και recurrent neural networks αποτέλεσαν τις πρώτες αρχιτεκτονικές που εφαρμόστηκαν στο TTS, επιτρέποντας την εκμετάλλευση τόσο τοπικών όσο και χρονικών συσχετίσεων στα δεδομένα (Mehri et al., 2023).

Στη συνέχεια, η εμφάνιση των end-to-end μοντέλων, όπως το Tacotron και οι παραλλαγές του, εισήγαγε μια ενιαία αρχιτεκτονική που συνδυάζει τον acoustic modeling και την ευθυγράμμιση σε ένα ενιαίο δίκτυο. Η χρήση μηχανισμών attention επέτρεψε τη δυναμική αντιστοίχιση μεταξύ εισόδου και εξόδου, εξαλείφοντας την ανάγκη για εξωτερικά εργαλεία ευθυγράμμισης. Παράλληλα, η εισαγωγή monotonic alignment loss σε επεκτάσεις όπως το Regotron βελτίωσε τη σταθερότητα της εκπαίδευσης και μείωσε τα φαινόμενα αποσυγχρονισμού (Georgiou et al., 2023). Επιπλέον, η συνδυαστική χρήση neural vocoders, όπως το WaveGlow, επέτρεψε την παραγωγή υψηλής ποιότητας κυματομορφών απευθείας από φασματικές αναπαραστάσεις, μειώνοντας σημαντικά τον θόρυβο και τις παραμορφώσεις (Rai et al., 2024). Ακολούθως, η εξέλιξη προς non-autoregressive μοντέλα αποτέλεσε καθοριστικό βήμα για τη βελτίωση της υπολογιστικής αποδοτικότητας. Σε αντίθεση με τα autoregressive μοντέλα, τα οποία παράγουν την έξοδο διαδοχικά, τα non-autoregressive συστήματα επιτρέπουν την παράλληλη παραγωγή των φωνητικών χαρακτηριστικών, μειώνοντας τον χρόνο inference έως και κατά μία τάξη μεγέθους. Μοντέλα όπως το FastSpeech και οι παραλλαγές του επιτυγχάνουν real-time factors μικρότερους του 0.1, καθιστώντας τα κατάλληλα για εφαρμογές πραγματικού χρόνου (Gao et al., 2024). Επιπλέον, η ενσωμάτωση duration predictors επιτρέπει τον έλεγχο της χρονικής δομής της ομιλίας, βελτιώνοντας την ακρίβεια της ευθυγράμμισης.

Η εισαγωγή transformer-based αρχιτεκτονικών επέφερε σημαντικές βελτιώσεις στη μοντελοποίηση μακροπρόθεσμων εξαρτήσεων και στη σημασιολογική κατανόηση του

κειμένου. Οι transformers, αξιοποιώντας μηχανισμούς self-attention, επιτρέπουν την επεξεργασία ολόκληρων ακολουθιών ταυτόχρονα, γεγονός που ενισχύει τη συνοχή και τη φυσικότητα της παραγόμενης ομιλίας (Latif et al., 2023). Επιπλέον, μοντέλα όπως το SpeechT5 και το Clip-TTS ενσωματώνουν πολυτροπικές πληροφορίες, επιτρέποντας τη βελτιωμένη συσχέτιση μεταξύ κειμένου και ακουστικών χαρακτηριστικών (Liu, 2025). Επιπροσθέτως, η εμφάνιση των diffusion-based μοντέλων εισήγαγε ένα νέο παράδειγμα στη σύνθεση φωνής, βασισμένο σε στοχαστικές διαδικασίες που προσομοιώνουν τη σταδιακή αποθορυβοποίηση ενός σήματος. Τα μοντέλα αυτά επιτυγχάνουν υψηλή πιστότητα και εκφραστικότητα, ενώ παράλληλα προσφέρουν μεγαλύτερο έλεγχο στη διαδικασία παραγωγής. Η χρήση style diffusion επιτρέπει τη μοντελοποίηση συναισθηματικών και προσωδιακών χαρακτηριστικών, οδηγώντας σε πιο φυσική και εκφραστική ομιλία (Kishore, 2025). Παράλληλα, η ενσωμάτωση σημασιολογικής πληροφορίας μέσω μοντέλων όπως το Llama-VITS βελτιώνει τη συνοχή μεταξύ κειμένου και παραγόμενης φωνής (Feng & Yoshimoto, 2024).

Η χρήση προεκπαιδευμένων μοντέλων και τεχνικών μεταφοράς γνώσης έχει μειώσει σημαντικά το κόστος ανάπτυξης TTS συστημάτων. Η δυνατότητα fine-tuning σε μικρά datasets επιτρέπει την προσαρμογή σε νέα γλωσσικά ή φωνητικά περιβάλλοντα με περιορισμένο υπολογιστικό κόστος, ενώ η χρήση synthetic data επιταχύνει τη διαδικασία εκπαίδευσης και βελτιώνει τη γενίκευση των μοντέλων (Liu, 2023; Joshi & Garera, 2023). Επιπλέον, η αξιοποίηση ενοποιημένων αρχιτεκτονικών όπως το VITS μειώνει την ανάγκη για ξεχωριστά υπομοντέλα, περιορίζοντας την πολυπλοκότητα του συστήματος (Aghalarov et al., 2025). Εξίσου σημαντική είναι η πρόοδος στη μοντελοποίηση της εκφραστικότητας και της προσωδίας, η οποία αποτελεί κρίσιμο παράγοντα για την αποδοχή των TTS συστημάτων από τους χρήστες. Η ενσωμάτωση latent representations και speaker embeddings επιτρέπει την παραγωγή φωνών με διαφορετικά χαρακτηριστικά, ενώ τεχνικές όπως StyleTTS επιτρέπουν την ελεγχόμενη μεταβολή του ύφους και του συναισθήματος (Han & Mesgarani, 2022). Παράλληλα, η χρήση prosody modeling έχει αποδειχθεί καθοριστική για τη βελτίωση της φυσικότητας, καθώς η προσωδία επηρεάζει άμεσα την αντίληψη της ομιλίας (Galdino et al., 2025).

Ταυτόχρονα, η πρόοδος στη συμπίεση και αναπαράσταση της ομιλίας μέσω τεχνικών όπως speech tokenization και vector quantization έχει συμβάλει στη βελτίωση της αποδοτικότητας των μοντέλων. Οι προσεγγίσεις αυτές επιτρέπουν την αναπαράσταση της ομιλίας σε διακριτά σύμβολα, μειώνοντας τον όγκο των δεδομένων και επιταχύνοντας τη διαδικασία εκπαίδευσης (Bai et al., 2024; Chen et al., 2023). Επιπλέον, η χρήση latent speech representations, όπως στο WavThruVec, επιτρέπει την αποτελεσματικότερη μεταφορά πληροφορίας μεταξύ διαφορετικών σταδίων του pipeline (Siuzdak et al., 2022). Η εφαρμογή των νευρωνικών TTS συστημάτων συνοδεύεται από νέες προκλήσεις που σχετίζονται με την ασφάλεια, την ιδιωτικότητα και την ηθική χρήση της τεχνολογίας. Η δυνατότητα δημιουργίας ρεαλιστικών συνθετικών φωνών εγείρει ζητήματα κατάχρησης, καθιστώντας απαραίτητη την ανάπτυξη μηχανισμών ανίχνευσης και προστασίας (Zhang et al., 2023). Παράλληλα, η ανάπτυξη datasets για την ανίχνευση συνθετικής ομιλίας υπογραμμίζει τη σημασία της υπεύθυνης χρήσης της τεχνολογίας (Salvi et al., 2023).

2.4 Εισαγωγή στα Προεκπαιδευμένα Μοντέλα

Η έννοια των προεκπαιδευμένων μοντέλων (pretrained models) έχει αναδειχθεί τα τελευταία χρόνια ως θεμελιώδης συνιστώσα της σύγχρονης μηχανικής μάθησης, μετασχηματίζοντας ριζικά τον τρόπο με τον οποίο σχεδιάζονται, εκπαιδεύονται και αναπτύσσονται συστήματα σύνθεσης φωνής. Υπό αυτή την οπτική, η μετάβαση από την εκπαίδευση μοντέλων από μηδενική βάση (training from scratch) προς την αξιοποίηση προεκπαιδευμένων αρχιτεκτονικών συνιστά μια στρατηγική επιλογή υψηλής αποδοτικότητας, καθώς επιτρέπει την επαναχρησιμοποίηση γνώσης που έχει εξαχθεί από μεγάλης κλίμακας δεδομένα. Στο πεδίο της σύνθεσης φωνής, η πρακτική αυτή αποκτά ιδιαίτερη σημασία, δεδομένου ότι η συλλογή και επισημείωση ηχητικών δεδομένων υψηλής ποιότητας αποτελεί μια ιδιαίτερα δαπανηρή και χρονοβόρα διαδικασία (Liu, 2023).

Ειδικότερα, τα προεκπαιδευμένα μοντέλα βασίζονται στην αρχή της μεταφοράς γνώσης (transfer learning), σύμφωνα με την οποία ένα μοντέλο εκπαιδεύεται αρχικά σε ένα μεγάλο και γενικευμένο σύνολο δεδομένων και στη συνέχεια προσαρμόζεται σε μια συγκεκριμένη εργασία μέσω διαδικασιών fine-tuning. Η προσέγγιση αυτή

επιτρέπει την εκμάθηση γενικών αναπαραστάσεων της ομιλίας, οι οποίες μπορούν να επαναχρησιμοποιηθούν σε διαφορετικά γλωσσικά και φωνητικά περιβάλλοντα. Στην περίπτωση των TTS συστημάτων, η χρήση προεκπαιδευμένων μοντέλων έχει αποδειχθεί ιδιαίτερα αποτελεσματική σε σενάρια few-shot learning, όπου η διαθέσιμη ποσότητα δεδομένων είναι περιορισμένη, επιτυγχάνοντας ικανοποιητικά επίπεδα φυσικότητας με ελάχιστα λεπτά ηχογράφησης (Liu, 2023).

Η χρήση latent representations επιτρέπει τη συμπίεση της πληροφορίας σε διανυσματικούς χώρους χαμηλότερης διάστασης, διευκολύνοντας τη μεταφορά γνώσης μεταξύ διαφορετικών εργασιών. Επιπλέον, τεχνικές όπως η vector quantization και η speech tokenization επιτρέπουν τη διακριτοποίηση της ομιλίας σε σύμβολα, καθιστώντας δυνατή την εφαρμογή μεθόδων που έχουν αναπτυχθεί αρχικά για την επεξεργασία φυσικής γλώσσας (Chen et al., 2023; Bai et al., 2024). Από διαφορετική σκοπιά, η αξιοποίηση προεκπαιδευμένων μοντέλων έχει επιφέρει σημαντικές βελτιώσεις στην υπολογιστική αποδοτικότητα των TTS συστημάτων. Συγκεκριμένα, η ανάγκη για εκπαίδευση μεγάλων μοντέλων από την αρχή, η οποία μπορεί να απαιτεί εκατοντάδες ώρες GPU και σημαντικούς ενεργειακούς πόρους, αντικαθίσταται από διαδικασίες fine-tuning που ολοκληρώνονται σε πολύ μικρότερο χρονικό διάστημα. Η μείωση αυτή του υπολογιστικού κόστους καθιστά τη σύνθεση φωνής προσβάσιμη σε ευρύτερο φάσμα ερευνητών και επιχειρήσεων, επιταχύνοντας την καινοτομία στον τομέα (Ma et al., 2025).

Η χρήση προεκπαιδευμένων μοντέλων επιτρέπει την ανάπτυξη πολυγλωσσικών και πολυφωνικών συστημάτων, τα οποία μπορούν να προσαρμοστούν σε διαφορετικές γλώσσες και ομιλητές χωρίς την ανάγκη εκτεταμένης επανεκπαίδευσης. Η δυνατότητα αυτή είναι ιδιαίτερα σημαντική για γλώσσες χαμηλών πόρων, όπου η διαθεσιμότητα δεδομένων είναι περιορισμένη. Μέσω της μεταφοράς γνώσης από γλώσσες υψηλών πόρων, τα μοντέλα μπορούν να επιτύχουν αποδεκτά επίπεδα απόδοσης ακόμη και σε περιβάλλοντα με ελάχιστα δεδομένα (Katumba et al., 2025). Παράλληλα, η ενσωμάτωση τεχνικών semi-supervised και self-supervised μάθησης επιτρέπει την αξιοποίηση μη επισημασμένων δεδομένων, ενισχύοντας περαιτέρω την αποδοτικότητα της εκπαίδευσης (Tu et al., 2020). Ακολούθως, η εξέλιξη των προεκπαιδευμένων μοντέλων στη σύνθεση φωνής συνδέεται στενά με την ανάπτυξη transformer-based αρχιτεκτονικών, οι οποίες έχουν αποδειχθεί ιδιαίτερα αποτελεσματικές στην

επεξεργασία ακολουθιών και στη μοντελοποίηση σύνθετων εξαρτήσεων. Τα μοντέλα αυτά αξιοποιούν μηχανισμούς self-attention για την εκμάθηση σχέσεων μεταξύ διαφορετικών στοιχείων της εισόδου, επιτρέποντας τη δημιουργία πιο συνεκτικής και φυσικής ομιλίας (Latif et al., 2023). Επιπλέον, η ενσωμάτωση πολυτροπικής πληροφορίας, όπως στο Clip-TTS, επιτρέπει τη βελτιωμένη συσχέτιση μεταξύ κειμένου και ακουστικών χαρακτηριστικών, οδηγώντας σε πιο ακριβή και εκφραστική σύνθεση φωνής (Liu, 2025).

Η δε ανάπτυξη ενοποιημένων αρχιτεκτονικών, όπως το VITS, επιτρέπει τη συνδυασμένη εκμάθηση acoustic modeling και waveform generation σε ένα ενιαίο μοντέλο, μειώνοντας την πολυπλοκότητα του pipeline και βελτιώνοντας την αποδοτικότητα (Aghalarov et al., 2025). Η προσέγγιση αυτή, σε συνδυασμό με τη χρήση προεκπαιδευμένων παραμέτρων, επιτρέπει την ταχεία ανάπτυξη συστημάτων υψηλής ποιότητας, χωρίς την ανάγκη για εκτεταμένη ρύθμιση υπερπαραμέτρων. Εξίσου σημαντική είναι η συμβολή των προεκπαιδευμένων μοντέλων στην ενίσχυση της εκφραστικότητας της συνθετικής ομιλίας. Μέσω της εκμάθησης πλούσιων αναπαραστάσεων που περιλαμβάνουν πληροφορία σχετικά με το ύφος, το συναίσθημα και την προσωδία, τα μοντέλα αυτά μπορούν να παράγουν ομιλία που δεν περιορίζεται σε ουδέτερες εκφωνήσεις, αλλά προσεγγίζει την ποικιλία της ανθρώπινης έκφρασης. Τεχνικές όπως το style transfer και η χρήση speaker embeddings επιτρέπουν την ελεγχόμενη μεταβολή των χαρακτηριστικών της φωνής, διευρύνοντας το φάσμα των εφαρμογών (Han & Mesgarani, 2022).

Η μείωση των απαιτήσεων σε δεδομένα και υπολογιστικούς πόρους οδηγεί σε χαμηλότερο κόστος παραγωγής, ενώ η δυνατότητα επαναχρησιμοποίησης μοντέλων επιτρέπει την ταχύτερη είσοδο στην αγορά. Επιπλέον, η χρήση cloud-based πλατφορμών, όπως το Hugging Face, διευκολύνει την πρόσβαση σε προεκπαιδευμένα μοντέλα και εργαλεία, μειώνοντας το τεχνολογικό φράγμα για νέους χρήστες. Τέλος, παρά τα σημαντικά πλεονεκτήματα, η χρήση προεκπαιδευμένων μοντέλων συνοδεύεται από προκλήσεις που σχετίζονται με τη γενίκευση, την ασφάλεια και την ηθική χρήση της τεχνολογίας. Η εξάρτηση από δεδομένα εκπαίδευσης μεγάλης κλίμακας ενδέχεται να οδηγήσει σε μεροληψίες, ενώ η δυνατότητα δημιουργίας ρεαλιστικών συνθετικών φωνών εγείρει ζητήματα κατάχρησης και προστασίας προσωπικών δεδομένων (Zhang et al., 2023). Επιπλέον, η ανάγκη για αξιόπιστη

αξιολόγηση της ποιότητας και της αυθεντικότητας της ομιλίας καθιστά απαραίτητη την ανάπτυξη νέων μεθοδολογιών και εργαλείων.

2.5 Η πλατφόρμα Hugging Face και το οικοσύστημά της

Η ανάπτυξη της πλατφόρμας Hugging Face ως κεντρικού κόμβου ανάπτυξης και διάθεσης μοντέλων μηχανικής μάθησης συνιστά μία από τις πλέον καθοριστικές εξελίξεις στο σύγχρονο οικοσύστημα της τεχνητής νοημοσύνης, επηρεάζοντας άμεσα και το πεδίο της σύνθεσης φωνής. Υπό αυτή την έννοια, η Hugging Face δεν αποτελεί απλώς μια αποθήκη προεκπαιδευμένων μοντέλων, αλλά ένα ολοκληρωμένο οικοσύστημα που ενοποιεί δεδομένα, αλγορίθμους και εργαλεία ανάπτυξης, επιτρέποντας την ταχεία υλοποίηση και προσαρμογή πολύπλοκων συστημάτων TTS. Η φιλοσοφία της πλατφόρμας βασίζεται στην ανοιχτή επιστήμη και στη συνεργατική ανάπτυξη, γεγονός που έχει οδηγήσει στη δημιουργία ενός αποθετηρίου με χιλιάδες μοντέλα και datasets, τα οποία μπορούν να αξιοποιηθούν άμεσα από την ερευνητική και βιομηχανική κοινότητα (Ma et al., 2025).

Η βιβλιοθήκη Transformers, η οποία παρέχει ένα ενοποιημένο περιβάλλον για την υλοποίηση και χρήση μοντέλων βασισμένων σε αρχιτεκτονικές attention. Αν και αρχικά σχεδιάστηκε για εφαρμογές επεξεργασίας φυσικής γλώσσας, η βιβλιοθήκη αυτή έχει επεκταθεί ώστε να υποστηρίζει και μοντέλα επεξεργασίας ομιλίας, επιτρέποντας την αξιοποίηση κοινών αρχιτεκτονικών για πολλαπλές εργασίες. Η δυνατότητα αυτή είναι ιδιαίτερα σημαντική για τη σύνθεση φωνής, καθώς επιτρέπει τη μεταφορά γνώσης μεταξύ διαφορετικών τομέων, ενισχύοντας την απόδοση των μοντέλων (Latif et al., 2023). Επιπλέον, η χρήση self-attention μηχανισμών διευκολύνει τη μοντελοποίηση μακροπρόθεσμων εξαρτήσεων, οδηγώντας σε πιο συνεκτική και φυσική ομιλία.

Η βιβλιοθήκη Datasets αποτελεί έναν δεύτερο κρίσιμο πυλώνα του οικοσυστήματος, παρέχοντας πρόσβαση σε μεγάλης κλίμακας σύνολα δεδομένων, τα οποία είναι απαραίτητα για την εκπαίδευση και αξιολόγηση μοντέλων TTS. Η δυνατότητα εύκολης φόρτωσης, επεξεργασίας και διαχείρισης δεδομένων μειώνει σημαντικά τον χρόνο ανάπτυξης, ενώ παράλληλα εξασφαλίζει τη συμβατότητα μεταξύ διαφορετικών πειραματικών ρυθμίσεων. Επιπροσθέτως, η υποστήριξη streaming δεδομένων επιτρέπει την εκπαίδευση μοντέλων σε μεγάλα datasets χωρίς την ανάγκη

αποθήκευσης τους τοπικά, μειώνοντας τις απαιτήσεις σε αποθηκευτικό χώρο (Yiwere et al., 2022). Ακολούθως, η πλατφόρμα Hugging Face ενσωματώνει το Hugging Face Hub, το οποίο λειτουργεί ως κεντρικό αποθετήριο μοντέλων και datasets. Μέσω του Hub, οι χρήστες μπορούν να αναζητήσουν, να κατεβάσουν και να επαναχρησιμοποιήσουν προεκπαιδευμένα μοντέλα, ενώ παράλληλα έχουν τη δυνατότητα να δημοσιεύσουν τα δικά τους μοντέλα, συμβάλλοντας στην ανάπτυξη της κοινότητας. Η πρακτική αυτή ενισχύει τη διαφάνεια και την αναπαραγωγιμότητα της έρευνας, καθώς επιτρέπει την εύκολη πρόσβαση σε πειραματικά αποτελέσματα και παραμέτρους εκπαίδευσης (Kharitonov et al., 2023).

Η πλατφόρμα Spaces παρέχει ένα περιβάλλον ανάπτυξης και φιλοξενίας εφαρμογών μηχανικής μάθησης, επιτρέποντας την άμεση αλληλεπίδραση με μοντέλα μέσω γραφικών διεπαφών. Η δυνατότητα αυτή είναι ιδιαίτερα χρήσιμη για την αξιολόγηση συστημάτων σύνθεσης φωνής, καθώς επιτρέπει τη γρήγορη δοκιμή διαφορετικών μοντέλων και παραμέτρων από μη εξειδικευμένους χρήστες. Παράλληλα, η ενσωμάτωση εργαλείων όπως το Gradio και το Streamlit διευκολύνει τη δημιουργία διαδραστικών εφαρμογών, μειώνοντας το τεχνικό φράγμα για την ανάπτυξη πρωτοτύπων.

Η δυνατότητα fine-tuning προεκπαιδευμένων μοντέλων σε μικρά datasets επιτρέπει την ανάπτυξη εξειδικευμένων εφαρμογών χωρίς την ανάγκη εκπαίδευσης από μηδενική βάση, μειώνοντας σημαντικά τις απαιτήσεις σε υπολογιστικούς πόρους (Liu, 2023). Επιπλέον, η υποστήριξη cloud-based υποδομών επιτρέπει την εκτέλεση πειραμάτων σε απομακρυσμένα περιβάλλοντα, ενισχύοντας την ευελιξία και την επεκτασιμότητα των συστημάτων. Εξίσου σημαντική είναι η συμβολή της πλατφόρμας στην επιτάχυνση της έρευνας και της καινοτομίας, καθώς παρέχει ένα κοινό σημείο αναφοράς για την ανταλλαγή γνώσης και τη σύγκριση διαφορετικών προσεγγίσεων. Η ύπαρξη standardized pipelines και evaluation metrics διευκολύνει τη συγκριτική ανάλυση μοντέλων, ενώ η πρόσβαση σε κοινά datasets επιτρέπει την αντικειμενική αξιολόγηση της απόδοσης (Galdino et al., 2025). Παράλληλα, η δυνατότητα ενσωμάτωσης νέων αρχιτεκτονικών και τεχνικών επιτρέπει την ταχεία υιοθέτηση καινοτόμων προσεγγίσεων, όπως τα diffusion-based μοντέλα και οι πολυτροπικές αρχιτεκτονικές.

Από διαφορετική σκοπιά, η Hugging Face διαδραματίζει σημαντικό ρόλο και στην υποστήριξη πολυγλωσσικών εφαρμογών, παρέχοντας πρόσβαση σε datasets και μοντέλα για ένα ευρύ φάσμα γλωσσών. Η δυνατότητα αυτή είναι ιδιαίτερα σημαντική για τη σύνθεση φωνής, καθώς επιτρέπει την ανάπτυξη συστημάτων για γλώσσες χαμηλών πόρων, αξιοποιώντας τεχνικές μεταφοράς γνώσης (Katumba et al., 2025). Επιπλέον, η ενσωμάτωση εργαλείων για την επεξεργασία και κανονικοποίηση κειμένου διευκολύνει την προσαρμογή των μοντέλων σε διαφορετικά γλωσσικά περιβάλλοντα. Η ευκολία πρόσβασης σε προεκπαιδευμένα μοντέλα υψηλής ποιότητας ενδέχεται να οδηγήσει σε καταχρήσεις, όπως η δημιουργία παραπλανητικού ή κακόβουλου περιεχομένου. Για τον λόγο αυτό, η ανάπτυξη μηχανισμών ανίχνευσης και προστασίας καθίσταται απαραίτητη, ιδιαίτερα σε εφαρμογές που αφορούν τη σύνθεση φωνής (Zhang et al., 2023). Παράλληλα, η ανάγκη για διασφάλιση της ιδιωτικότητας των δεδομένων υπογραμμίζει τη σημασία της υπεύθυνης χρήσης της τεχνολογίας.

Κεφάλαιο 3: Τεχνολογίες και Εργαλεία

3.1 Προεκπαιδευμένα Μοντέλα για Σύνθεση Φωνής

Η εντυπωσιακή επιτάχυνση της έρευνας στον τομέα της νευρωνικής σύνθεσης φωνής κατά την τελευταία πενταετία συνδέεται άμεσα με τη μετάβαση από απομονωμένες αρχιτεκτονικές εκπαίδευσης σε εκτεταμένα προεκπαιδευμένα μοντέλα βαθιάς μάθησης, τα οποία είναι ικανά να ενσωματώνουν φωνητική, σημασιολογική και προσωδιακή πληροφορία σε λανθάνουσες αναπαραστάσεις υψηλής διάστασης. Στο σύγχρονο τεχνολογικό περιβάλλον, τα pretrained TTS μοντέλα δεν λειτουργούν απλώς ως επαναχρησιμοποιήσιμα βάρη νευρωνικών δικτύων, αλλά ως πλήρεις μηχανισμοί μεταφοράς γνώσης που μειώνουν δραστικά το υπολογιστικό κόστος ανάπτυξης και επιταχύνουν την παραγωγή συστημάτων υψηλής πιστότητας. Σύμφωνα με τον Liu (2023), η εκπαίδευση ενός end-to-end TTS μοντέλου από μηδενική βάση σε dataset άνω των 20 ωρών μπορεί να υπερβεί τις 300 GPU-hours σε υποδομές NVIDIA A100, ενώ αντίστοιχες διαδικασίες fine-tuning σε προεκπαιδευμένες αρχιτεκτονικές απαιτούν λιγότερο από το 8% του συνολικού αυτού χρόνου. Η τεχνοοικονομική αυτή μεταβολή έχει επηρεάσει άμεσα τόσο την ακαδημαϊκή έρευνα όσο και τη βιομηχανική ανάπτυξη

εφαρμογών σύνθεσης φωνής, διότι επέτρεψε την είσοδο μικρότερων ερευνητικών ομάδων σε ένα πεδίο που προηγουμένως απαιτούσε ιδιαίτερα υψηλή υπολογιστική επένδυση.

Η μετάβαση προς pretrained νευρωνικές δομές συνδέθηκε με ουσιαστική βελτίωση της φυσικότητας της παραγόμενης ομιλίας. Ενώ τα παραδοσιακά parametric συστήματα σπάνια ξεπερνούσαν Mean Opinion Score (MOS) ίσο με 3.2–3.5, τα σύγχρονα νευρωνικά μοντέλα όπως Tacotron2, VITS και StyleTTS επιτυγχάνουν σταθερά τιμές άνω του 4.2 σε πενταβάθμια κλίμακα αξιολόγησης (Maniati et al., 2022). Το γεγονός αυτό δεν οφείλεται αποκλειστικά στην αύξηση της πολυπλοκότητας των αρχιτεκτονικών, αλλά κυρίως στη δυνατότητα των προεκπαιδευμένων μοντέλων να αποκτούν γενικευμένες αναπαραστάσεις φωνητικής πληροφορίας μέσω εκπαίδευσης σε μεγάλης κλίμακας corpora. Επιπροσθέτως, η χρήση transfer learning έχει επιτρέψει την ανάπτυξη συστημάτων σύνθεσης φωνής ακόμη και σε γλώσσες εξαιρετικά χαμηλών πόρων, όπου η διαθέσιμη ηχογράφηση περιορίζεται σε λιγότερο από μία ώρα δεδομένων, όπως καταγράφηκε στις εργασίες των Katumba et al. (2025) και Muhamad και Veisi (2022).

Σημασία απέκτησε το Tacotron2, το οποίο εξακολουθεί να θεωρείται ένα από τα πλέον επιδραστικά μοντέλα στη νευρωνική σύνθεση φωνής λόγω της αρχιτεκτονικής encoder–attention–decoder που εισήγαγε. Η δομή αυτή επιτρέπει τη μετατροπή ακολουθιών χαρακτήρων σε mel-spectrograms μέσω attention alignment, εξαλείφοντας την ανάγκη εξωτερικής ευθυγράμμισης μεταξύ κειμένου και φωνημάτων. Οι Rai et al. (2024) κατέδειξαν ότι Tacotron2 σε συνδυασμό με WaveGlow vocoder παρουσίασε μείωση Mel-Cepstral Distortion κάτω από 4.1 dB ακόμη και σε low-resource περιβάλλοντα, ενώ παράλληλα η παραγόμενη ομιλία εμφάνισε intelligibility scores μεγαλύτερα του 92%. Ωστόσο, παρά τη σημαντική ακουστική ποιότητα, η autoregressive φύση του Tacotron2 δημιουργεί αυξημένο latency κατά το inference, καθώς κάθε φασματικό frame παράγεται σειριακά. Σε πρακτικές συνθήκες deployment, αυτό μεταφράζεται σε real-time factor που συχνά κυμαίνεται μεταξύ 0.7 και 1.1, περιορίζοντας τη χρήση του σε εφαρμογές άμεσης απόκρισης.

Αντίθετα, οι νεότερες unified αρχιτεκτονικές όπως το VITS μετασχημάτισαν ριζικά τη φιλοσοφία σχεδιασμού TTS συστημάτων, ενοποιώντας acoustic model, variational

inference και waveform generation σε ενιαίο end-to-end πλαίσιο. Οι Aghalarov et al. (2025) υπογράμμισαν ότι η χρήση του VITS μειώνει έως και κατά 35% την πολυπλοκότητα ανάπτυξης συγκριτικά με pipelines πολλαπλών υπομοντέλων, ενώ ταυτόχρονα βελτιώνει την εκφραστικότητα και τη φυσικότητα της ομιλίας. Επιπλέον, η non-autoregressive λειτουργία του επιτρέπει massively parallel inference, μειώνοντας το latency σε επίπεδα κάτω των 80 milliseconds ανά δευτερόλεπτο συνθετικής ομιλίας. Η τεχνοοικονομική συνέπεια αυτής της εξέλιξης είναι ιδιαίτερα σημαντική, διότι η μείωση inference cost οδηγεί σε μικρότερη κατανάλωση GPU πόρων σε cloud υπηρεσίες μεγάλης κλίμακας. Σε περιβάλλοντα commercial deployment, η διαφορά αυτή μπορεί να μειώσει το λειτουργικό κόστος έως και 40% σε εφαρμογές με εκατομμύρια καθημερινά requests.

Ταυτόχρονα, η ενσωμάτωση transformer-based μηχανισμών σε προεκπαιδευμένα μοντέλα σύνθεσης φωνής βελτίωσε ουσιαστικά τη σημασιολογική συνοχή της παραγόμενης ομιλίας. Οι Latif et al. (2023) εξήγησαν ότι τα self-attention mechanisms επιτρέπουν τη μοντελοποίηση μακροπρόθεσμων εξαρτήσεων μεταξύ φωνητικών και γλωσσικών μονάδων, κάτι που ήταν εξαιρετικά δύσκολο στις recurrent αρχιτεκτονικές προηγούμενης γενιάς. Μέσα σε αυτό το πλαίσιο, μοντέλα όπως SpeechT5 και Clip-TTS αξιοποιούν contrastive learning και multimodal embeddings, ώστε να συσχετίζουν το κείμενο με φασματικές και σημασιολογικές ιδιότητες της ομιλίας. Ο Liu (2025) ανέφερε ότι το Clip-TTS παρουσίασε βελτίωση περίπου 18% στη semantic consistency συγκριτικά με συμβατικά transformer-TTS μοντέλα, επιτυγχάνοντας παράλληλα χαμηλότερα alignment errors σε μεγάλες προτάσεις άνω των 30 tokens.

Εξίσου καθοριστική υπήρξε η ανάπτυξη diffusion-based TTS μοντέλων, τα οποία εισήγαγαν στοχαστικές διαδικασίες αποθορυβοποίησης για την παραγωγή ηχητικών κυματομορφών εξαιρετικά υψηλής πιστότητας. Σε αντίθεση με τις deterministic προσεγγίσεις, τα diffusion models παράγουν την ομιλία μέσω διαδοχικών βημάτων refinement, επιτρέποντας ακριβέστερο έλεγχο προσωδίας, τόνου και συναισθηματικής χροιάς. Ο Kishore (2025) κατέγραψε ότι style diffusion architectures πέτυχαν MOS τιμές που άγγιζαν το 4.48, προσεγγίζοντας την ανθρώπινη ομιλία σε controlled listening tests. Παρά τα σημαντικά αυτά πλεονεκτήματα, η διαδικασία iterative denoising παραμένει ιδιαίτερα απαιτητική υπολογιστικά, καθώς ορισμένα diffusion pipelines απαιτούν άνω των 50 sampling steps ανά εκφώνηση. Αυτό συνεπάγεται

υψηλότερη κατανάλωση ενέργειας και σημαντικά αυξημένο inference cost σε σύγκριση με non-autoregressive architectures όπως FastSpeech ή VITS.

Η δε πρόοδος στα pretrained speech representations επέτρεψε τη χρήση vector quantization και speech tokenization ως μηχανισμών συμπίεσης της ομιλίας σε διακριτές μονάδες. Οι Chen et al. (2023) απέδειξαν ότι vector quantized latent representations μπορούν να μειώσουν έως και κατά 60% τον όγκο των δεδομένων χωρίς αισθητή απώλεια ακουστικής ποιότητας. Παράλληλα, οι Bai et al. (2024) παρουσίασαν το dMel framework, στο οποίο η speech tokenization απλοποιεί σημαντικά τη διαδικασία εκπαίδευσης και επιτρέπει ταχύτερη σύγκλιση των μοντέλων. Η οικονομική σημασία αυτών των τεχνικών είναι εξαιρετικά μεγάλη, καθώς η μείωση αποθηκευτικού και μεταφορικού κόστους επιτρέπει την ανάπτυξη edge-based εφαρμογών σύνθεσης φωνής σε συσκευές περιορισμένων πόρων. Οι Mbonimpa et al. (2025) επεσήμαναν ότι edge-based TTS συστήματα για Kinyarwanda και Swahili κατάφεραν να λειτουργήσουν σε low-power embedded hardware με latency κάτω των 150 milliseconds, χωρίς ανάγκη συνεχούς cloud επικοινωνίας.

Από διαφορετική σκοπιά, η χρήση προεκπαιδευμένων μοντέλων για speaker adaptation και voice cloning έχει δημιουργήσει ένα νέο επιχειρηματικό οικοσύστημα γύρω από προσωποποιημένες υπηρεσίες σύνθεσης φωνής. Οι Joshi και Garera (2023) απέδειξαν ότι synthetic augmentation techniques μπορούν να επιταχύνουν έως και κατά 40% τη σύγκλιση μοντέλων speaker adaptation, ενώ οι Das et al. (2025) έδειξαν ότι single-speaker voice cloning μέσω BnVITS μπορεί να επιτύχει υψηλή ομοιότητα φωνής με λιγότερο από 15 λεπτά ηχογραφημένου υλικού. Παρότι οι τεχνολογίες αυτές ανοίγουν νέες δυνατότητες στην προσωποποιημένη αλληλεπίδραση ανθρώπου-μηχανής, δημιουργούν ταυτόχρονα σοβαρές προκλήσεις ασφαλείας. Οι Zhang et al. (2023) κατέγραψαν αυξημένο κίνδυνο μη εξουσιοδοτημένης αναπαραγωγής φωνής και υπογράμμισαν την ανάγκη ανάπτυξης anti-spoofing μηχανισμών προστασίας. Παράλληλα, οι Salvi et al. (2023) ανέπτυξαν το dataset TIMIT-TTS για την ανίχνευση συνθετικής ομιλίας, επιβεβαιώνοντας ότι η πρόοδος στην ποιότητα TTS συστημάτων καθιστά ολοένα δυσκολότερη τη διάκριση μεταξύ ανθρώπινης και συνθετικής φωνής.

3.2 Βιβλιοθήκες Hugging Face (Transformers, Datasets, Spaces)

Η εντυπωσιακή εξάπλωση των προεκπαιδευμένων μοντέλων βαθιάς μάθησης στον χώρο της σύνθεσης φωνής συνδέεται άμεσα με την ανάδυση ολοκληρωμένων οικοσυστημάτων ανάπτυξης, τα οποία επιτρέπουν τη διαχείριση δεδομένων, την εκπαίδευση αρχιτεκτονικών υψηλής πολυπλοκότητας και τη διαμοίραση μοντέλων σε μεγάλη κλίμακα. Υπό αυτό το πρίσμα, το οικοσύστημα της Hugging Face αποτελεί σήμερα έναν από τους πλέον καθοριστικούς τεχνολογικούς πυλώνες της σύγχρονης υπολογιστικής γλωσσολογίας και των speech technologies, καθώς συγκεντρώνει σε ενιαία υποδομή εργαλεία εκπαίδευσης, inference, fine-tuning και deployment νευρωνικών μοντέλων. Η στρατηγική σημασία της πλατφόρμας δεν περιορίζεται αποκλειστικά στη διευκόλυνση της έρευνας, αλλά επεκτείνεται και στη δραστική μείωση του τεχνοοικονομικού κόστους ανάπτυξης εφαρμογών τεχνητής νοημοσύνης. Σύμφωνα με τους Ma et al. (2025), η χρήση reusable pretrained architectures μέσω Hugging Face μειώνει κατά 55–70% τον χρόνο αρχικής ανάπτυξης speech εφαρμογών συγκριτικά με παραδοσιακές custom implementations, ενώ ταυτόχρονα περιορίζει σημαντικά τις απαιτήσεις σε computational engineering manpower.

Κεντρικό ρόλο στο οικοσύστημα αυτό διαδραματίζει η βιβλιοθήκη Transformers, η οποία αποτελεί ουσιαστικά το βασικό abstraction layer για την υλοποίηση attention-based νευρωνικών αρχιτεκτονικών. Η συγκεκριμένη βιβλιοθήκη, παρότι αρχικά σχεδιάστηκε για Natural Language Processing εφαρμογές, εξελίχθηκε ταχύτατα σε γενικευμένο framework multimodal machine learning, υποστηρίζοντας πλέον speech synthesis, speech recognition και audio-language models. Η αρχιτεκτονική των Transformers βασίζεται στον μηχανισμό self-attention, ο οποίος επιτρέπει τη μοντελοποίηση μακροπρόθεσμων εξαρτήσεων χωρίς την ανάγκη επαναληπτικών recurrent δομών. Οι Latif et al. (2023) επεσήμαναν ότι η computational complexity των transformer layers αυξάνεται θεωρητικά με τάξη $O(n^2)$ ως προς το sequence length, ωστόσο η massively parallel processing δυνατότητα των GPUs καθιστά τα transformers σημαντικά αποδοτικότερα από recurrent μοντέλα σε μεγάλης κλίμακας training workloads. Η πρακτική σημασία της βιβλιοθήκης Transformers στη σύνθεση φωνής γίνεται ιδιαίτερα εμφανής μέσω μοντέλων όπως SpeechT5, VITS, Whisper και Bark, τα οποία διατίθενται έτοιμα προς χρήση μέσα από standardized APIs. Επιπλέον, η δυνατότητα loading pretrained weights μέσω ελάχιστων γραμμών κώδικα μεταβάλλει

ριζικά τη διαδικασία ανάπτυξης TTS εφαρμογών. Ενδεικτικά, η φόρτωση ενός speech synthesis μοντέλου πραγματοποιείται με μία μόνο εντολή:

```
from transformers import SpeechT5Processor, SpeechT5ForTextToSpeech
```

Η απλοποίηση αυτή έχει ουσιαστικές τεχνοοικονομικές επιπτώσεις, διότι μειώνει δραστικά τον χρόνο integration σε enterprise περιβάλλοντα. Σε μεγάλης κλίμακας οργανισμούς, όπου το κόστος software engineering μπορεί να υπερβαίνει τα 120–180 ευρώ ανά ανθρωπόωρα, η μείωση development overhead οδηγεί σε σημαντική εξοικονόμηση πόρων κατά τη διάρκεια του lifecycle ενός speech προϊόντος.

Η Hugging Face παρέχει πλήρη συμβατότητα με PyTorch, TensorFlow και JAX, γεγονός που επιτρέπει την αξιοποίηση heterogeneous υπολογιστικών υποδομών. Οι Gao et al. (2024) υπογράμμισαν ότι η χρήση mixed precision training μέσω Hugging Face Accelerate μπορεί να μειώσει έως και κατά 42% την κατανάλωση VRAM σε diffusion-based TTS μοντέλα, επιτρέποντας την εκπαίδευση architectures άνω των 1.2 δισεκατομμυρίων παραμέτρων ακόμη και σε GPUs με 24 GB μνήμης. Η δυνατότητα αυτή αποκτά ιδιαίτερη σημασία σε ερευνητικά εργαστήρια περιορισμένου προϋπολογισμού, όπου η πρόσβαση σε high-end clusters παραμένει οικονομικά απαγορευτική.

Εξίσου κρίσιμος πυλώνας του οικοσυστήματος αποτελεί η βιβλιοθήκη Datasets, η οποία σχεδιάστηκε για τη διαχείριση μεγάλων συνόλων δεδομένων με υψηλή αποδοτικότητα. Η ανάπτυξη TTS συστημάτων εξαρτάται άμεσα από τη διαθεσιμότητα τεράστιων speech corpora, τα οποία συχνά υπερβαίνουν αρκετά terabytes σε μέγεθος. Οι Yiwere et al. (2022) ανέφεραν ότι πλήρη multi-speaker datasets υψηλής ποιότητας απαιτούν αποθηκευτικό χώρο άνω των 2.5 TB όταν περιλαμβάνονται raw waveforms, mel-spectrograms και alignment metadata. Μέσα σε αυτό το πλαίσιο, η βιβλιοθήκη Datasets εισήγαγε memory-mapped loading μηχανισμούς, οι οποίοι επιτρέπουν streaming δεδομένων χωρίς πλήρη φόρτωση στη RAM. Το χαρακτηριστικό αυτό μειώνει δραστικά τις απαιτήσεις συστήματος και επιτρέπει την εκπαίδευση μοντέλων σε commodity hardware. Ο ακόλουθος πίνακας παρουσιάζει συγκριτικά τις βασικές δυνατότητες των βιβλιοθηκών του οικοσυστήματος Hugging Face στη σύνθεση φωνής:

Βιβλιοθήκη	Βασική λειτουργία	Τεχνικά χαρακτηριστικά	Επίδραση στο TTS
Transformers	Υλοποίηση pretrained μοντέλων	Self-attention, GPU acceleration, mixed precision	Ταχύτερο training και inference
Datasets	Διαχείριση speech corpora	Streaming, memory mapping, lazy loading	Μείωση RAM και storage overhead
Spaces	Deployment εφαρμογών	Gradio, Streamlit, cloud hosting	Real-time demo και evaluation
Accelerate	Distributed training	Multi-GPU orchestration	Μείωση training χρόνου
Hub	Αποθετήριο μοντέλων	Versioning, model sharing	Αναπαραγωγιμότητα έρευνας

Πέραν της διαχείρισης δεδομένων, η Hugging Face εισήγαγε και μία ιδιαίτερα καινοτόμο προσέγγιση στη διαδικασία deployment μέσω της πλατφόρμας Spaces. Η συγκεκριμένη υπηρεσία λειτουργεί ως cloud-native περιβάλλον φιλοξενίας εφαρμογών machine learning, επιτρέποντας τη δημιουργία interactive demos χωρίς ανάγκη custom backend infrastructure. Η σημασία αυτής της δυνατότητας είναι εξαιρετικά μεγάλη στον χώρο της σύνθεσης φωνής, όπου η αξιολόγηση μοντέλων απαιτεί άμεση ακουστική αλληλεπίδραση. Μέσω εργαλείων όπως Gradio και Streamlit, οι ερευνητές μπορούν να δημιουργήσουν πλήρως λειτουργικά TTS interfaces σε ελάχιστο χρόνο. Ενδεικτικά, ένα βασικό deployment speech synthesis interface μπορεί να υλοποιηθεί ως εξής:

```
import gradio as gr

def synthesize(text):
    return model(text)

gr.Interface(fn=synthesize, inputs="text", outputs="audio").launch()
```

Η τεχνική απλοποίηση που προσφέρει το Spaces μεταφράζεται σε ουσιαστική επιχειρησιακή επιτάχυνση. Παραδοσιακά, η ανάπτυξη web-based inference interfaces απαιτούσε ξεχωριστή backend ομάδα, infrastructure engineers και DevOps

διαδικασίες. Αντιθέτως, μέσω Hugging Face Spaces, ακόμη και μικρές ερευνητικές ομάδες μπορούν να δημοσιεύσουν production-like demos μέσα σε λίγες ώρες. Οι Kharitonov et al. (2023) επεσήμαναν ότι η πρακτική αυτή επιτάχυνε σημαντικά την αναπαραγωγιμότητα πειραμάτων στον χώρο της σύνθεσης φωνής, καθώς τα μοντέλα συνοδεύονται πλέον από άμεσα προσβάσιμα interfaces αξιολόγησης. Από διαφορετική σκοπιά, η Hugging Face Hub λειτουργεί ως distributed repository pretrained μοντέλων, datasets και checkpoints. Μέχρι τις αρχές του 2025, η πλατφόρμα φιλοξενούσε περισσότερα από 1.2 εκατομμύρια μοντέλα και πάνω από 250 χιλιάδες datasets, με σημαντικό ποσοστό αυτών να αφορά speech technologies (Ma et al., 2025). Η δυνατότητα version control και model branching καθιστά εφικτή τη συνεχή βελτίωση αρχιτεκτονικών χωρίς απώλεια συμβατότητας μεταξύ διαφορετικών experimental pipelines.

Η δυνατότητα fine-tuning pretrained speech μοντέλων μέσω Hugging Face Trainer APIs. Οι Liu et al. (2022) έδειξαν ότι transfer learning pipelines βασισμένα σε pretrained TTS architectures επιτυγχάνουν convergence έως και 3.4 φορές ταχύτερα συγκριτικά με training from scratch προσεγγίσεις. Επιπλέον, η χρήση low-rank adaptation τεχνικών (LoRA) μειώνει δραστικά τον αριθμό trainable παραμέτρων, επιτρέποντας efficient adaptation ακόμη και σε περιβάλλοντα περιορισμένης GPU ισχύος. Ο ακόλουθος πίνακας συνοψίζει τις επιδράσεις του Hugging Face ecosystem στο κόστος ανάπτυξης TTS συστημάτων:

Παράμετρος	Παραδοσιακή ανάπτυξη	Hugging ecosystem	Face Μέση μείωση
Χρόνος αρχικού development	6–12 μήνες	1–3 μήνες	~70%
GPU training κόστος	>300 GPU-hours	<80 GPU-hours	~73%
Deployment overhead	Υψηλό	Χαμηλό	~60%
Ανάγκη custom infrastructure	Πολύ υψηλή	Περιορισμένη	~65%
Χρόνος experimentation	Εβδομάδες	Ημέρες	~75%

Περαιτέρω, η συμβολή της Hugging Face στη διάδοση multilingual TTS architectures είναι εξαιρετικά σημαντική. Οι Kumar et al. (2023) τόνισαν ότι η ύπαρξη shared multilingual pretrained models επιτρέπει την ανάπτυξη speech systems για «the next billion users», δηλαδή πληθυσμούς με περιορισμένη πρόσβαση σε computational infrastructure και speech datasets. Αντίστοιχα, οι Katumba et al. (2025) έδειξαν ότι crowdsourced δεδομένα μπορούν να αξιοποιηθούν αποτελεσματικά μέσω Hugging Face pipelines για low-resource αφρικανικές γλώσσες, επιτυγχάνοντας intelligibility scores άνω του 87% με datasets μικρότερα των 2 ωρών. Παρά τα σημαντικά αυτά πλεονεκτήματα, η αυξημένη προσβασιμότητα σε pretrained speech μοντέλα δημιουργεί νέους κινδύνους ασφαλείας. Οι Zhang et al. (2023) υπογράμμισαν ότι η εύκολη πρόσβαση σε voice cloning architectures μέσω public repositories αυξάνει τον κίνδυνο μη εξουσιοδοτημένης σύνθεσης φωνής. Παράλληλα, οι Salvi et al. (2023) επεσήμαναν ότι η συνεχής βελτίωση των TTS μοντέλων δυσχεραίνει σημαντικά την ανίχνευση synthetic speech, ιδιαίτερα όταν χρησιμοποιούνται diffusion-based architectures υψηλής πιστότητας.

3.3 Εργαλεία ανάπτυξης: Python, PyTorch/TensorFlow

Η θεαματική πρόοδος που σημειώθηκε τα τελευταία χρόνια στον χώρο της νευρωνικής σύνθεσης φωνής δεν μπορεί να αποσυνδεθεί από την εξέλιξη των σύγχρονων προγραμματιστικών και υπολογιστικών frameworks βαθιάς μάθησης, τα οποία επέτρεψαν την υλοποίηση μοντέλων με δισεκατομμύρια παραμέτρους και εξαιρετικά αυξημένες απαιτήσεις επεξεργαστικής ισχύος. Στο επίκεντρο αυτού του τεχνολογικού μετασχηματισμού βρίσκονται η γλώσσα προγραμματισμού Python και τα frameworks PyTorch και TensorFlow, τα οποία αποτελούν σήμερα τα κυρίαρχα εργαλεία ανάπτυξης συστημάτων Text-to-Speech (TTS). Η καθολική σχεδόν επικράτησή τους δεν σχετίζεται αποκλειστικά με την υποστήριξη νευρωνικών δικτύων, αλλά κυρίως με την ικανότητά τους να ενοποιούν διαδικασίες preprocessing, distributed training, automatic differentiation, GPU acceleration και deployment σε ένα ενιαίο οικοσύστημα ανάπτυξης. Σύμφωνα με τους Ma et al. (2025), ποσοστό μεγαλύτερο του 84% των σύγχρονων speech synthesis pipelines που δημοσιεύθηκαν μεταξύ 2022 και 2025 βασίστηκε είτε σε PyTorch είτε σε TensorFlow υποδομές, γεγονός που επιβεβαιώνει την τεχνολογική κυριαρχία των συγκεκριμένων frameworks.

Η Python κατέχει δεσπόζουσα θέση στον χώρο της τεχνητής νοημοσύνης λόγω της υψηλής αφαιρετικότητας, της ευρείας βιβλιοθήκης επιστημονικών εργαλείων και της εξαιρετικά χαμηλής καμπύλης εκμάθησης συγκριτικά με γλώσσες χαμηλότερου επιπέδου. Παρότι η εκτέλεση Python κώδικα χαρακτηρίζεται από χαμηλότερη ακατέργαστη υπολογιστική απόδοση σε σχέση με γλώσσες όπως η C++ ή η Rust, η δυνατότητα ενσωμάτωσης highly optimized backend libraries επιτρέπει την αποδοτική αξιοποίηση GPU clusters χωρίς ο προγραμματιστής να αλληλεπιδρά άμεσα με low-level CUDA kernels. Η πρακτική σημασία αυτού του χαρακτηριστικού είναι τεράστια για την ανάπτυξη TTS συστημάτων, διότι η εκπαίδευση μοντέλων σύνθεσης φωνής απαιτεί εξαιρετικά πολύπλοκους υπολογισμούς tensor algebra και massive parallel processing. Οι Tan et al. (2021) επεσήμαναν ότι η εκπαίδευση σύγχρονων neural vocoders μπορεί να περιλαμβάνει περισσότερες από 10^{15} floating point operations ανά training cycle, καθιστώντας αδύνατη τη χειροκίνητη διαχείριση computational graphs χωρίς εξειδικευμένα frameworks.

Σημαντικό πλεονέκτημα της Python αποτελεί επίσης η διαλειτουργικότητα με βιβλιοθήκες επεξεργασίας ήχου και επιστημονικών υπολογισμών. Εργαλεία όπως NumPy, SciPy, Librosa και SoundFile χρησιμοποιούνται εκτεταμένα για feature extraction, mel-spectrogram generation και audio normalization. Στο πλαίσιο της σύνθεσης φωνής, η μετατροπή ηχητικών κυματομορφών σε mel-frequency spectral representations αποτελεί κρίσιμο στάδιο preprocessing, καθώς επηρεάζει άμεσα την ποιότητα της τελικής παραγόμενης ομιλίας. Οι Hu et al. (2024) κατέδειξαν ότι βελτιωμένες τεχνικές mel-spectrogram enhancement μέσω Continuous Wavelet Transform (CWT) οδήγησαν σε αισθητηριακή βελτίωση άνω του 11% στην αντιληπτή καθαρότητα της ομιλίας, γεγονός που αναδεικνύει τη σημασία του signal processing επιπέδου πριν ακόμη την είσοδο των δεδομένων στο νευρωνικό μοντέλο.

Παράλληλα, το PyTorch εξελίχθηκε ταχύτατα στο κυρίαρχο framework έρευνας για speech synthesis applications, κυρίως λόγω της δυναμικής κατασκευής computational graphs και της ευελιξίας που παρέχει σε experimental αρχιτεκτονικές. Σε αντίθεση με παλαιότερες static graph προσεγγίσεις, το PyTorch επιτρέπει runtime graph execution, διευκολύνοντας την ανάπτυξη πολύπλοκων TTS architectures με adaptive alignment μηχανισμούς και variational latent spaces. Οι Georgiou et al. (2023) αξιοποίησαν αυτή τη δυναμική δυνατότητα στο Regotron framework, όπου monotonic alignment losses

εφαρμόζονται δυναμικά κατά τη διάρκεια της εκπαίδευσης για τη σταθεροποίηση του Tacotron2 alignment process. Ενδεικτικά, η δημιουργία tensor computations στο PyTorch πραγματοποιείται ως εξής:

```
import torch

x = torch.randn(16, 80, 400)
y = torch.nn.functional.relu(x)
```

Η απλότητα της σύνταξης αποκρύπτει στην πραγματικότητα μία ιδιαίτερα σύνθετη διαδικασία αυτόματης παραγωγίσης και GPU execution optimization. Η βιβλιοθήκη Autograd του PyTorch υπολογίζει αυτόματα gradients για δισεκατομμύρια παραμέτρους, επιτρέποντας efficient backpropagation ακόμη και σε extremely deep speech architectures. Οι Rai et al. (2024) υπογράμμισαν ότι η χρήση mixed precision training μέσω PyTorch CUDA AMP μείωσε έως και κατά 37% τον χρόνο εκπαίδευσης Tacotron2–WaveGlow pipelines χωρίς αισθητή απώλεια ακρίβειας.

Από τεχνοοικονομική σκοπιά, η μετάβαση σε PyTorch-based workflows επέφερε σημαντική μείωση development overhead. Παραδοσιακά, η ανάπτυξη custom GPU kernels απαιτούσε εξειδικευμένους CUDA engineers, αυξάνοντας κατακόρυφα το κόστος ανάπτυξης speech προϊόντων. Αντιθέτως, μέσω PyTorch abstractions, μικρές ομάδες μηχανικών μπορούν πλέον να αναπτύσσουν advanced TTS pipelines χωρίς εξειδικευμένη γνώση parallel computing architecture. Οι Gao et al. (2024) κατέγραψαν ότι η αξιοποίηση distributed PyTorch training σε multi-GPU clusters μείωσε κατά περίπου 63% τον συνολικό training χρόνο diffusion-based TTS μοντέλων συγκριτικά με single-device training configurations. Ο ακόλουθος πίνακας παρουσιάζει συγκριτικά βασικά χαρακτηριστικά των PyTorch και TensorFlow frameworks στη σύνθεση φωνής:

Χαρακτηριστικό	PyTorch	TensorFlow
Computational Graph	Dynamic	Static/Dynamic Hybrid
Ευκολία debugging	Πολύ υψηλή	Υψηλή
Research flexibility	Εξαιρετική	Πολύ καλή
Production deployment	Πολύ καλό	Εξαιρετικό
TPU support	Περιορισμένο	Πλήρες

Community adoption στο TTS	Πολύ υψηλή	Υψηλή
Distributed training efficiency	Υψηλή	Πολύ υψηλή

Αντίστοιχα, το TensorFlow εξακολουθεί να διατηρεί ιδιαίτερα ισχυρή παρουσία σε large-scale industrial deployments λόγω της βελτιστοποίησης που προσφέρει σε production inference περιβάλλοντα. Η υποστήριξη Tensor Processing Units (TPUs), η δυνατότητα graph optimization μέσω XLA compilation και η ενσωμάτωση TensorFlow Serving επιτρέπουν την ανάπτυξη speech synthesis υπηρεσιών με πολύ χαμηλό latency. Σε cloud-native περιβάλλοντα μεγάλης κλίμακας, το TensorFlow εμφανίζει σημαντικά πλεονεκτήματα όσον αφορά inference throughput και horizontal scalability. Οι Nguyen et al. (2022) έδειξαν ότι το Smartcall Vietnamese TTS system βασισμένο σε TensorFlow deployment infrastructure εξυπηρετούσε περισσότερα από 12.000 concurrent speech synthesis requests με μέσο latency κάτω των 95 milliseconds. Επιπλέον, η TensorFlow Lite υποδομή επιτρέπει τη μεταφορά TTS μοντέλων σε edge συσκευές περιορισμένων πόρων. Οι Mbonimpa et al. (2025) ανέφεραν ότι edge-based speech synthesis systems για Swahili και Kinyarwanda κατάφεραν να λειτουργήσουν σε embedded ARM hardware με κατανάλωση κάτω των 4W, αξιοποιώντας quantized TensorFlow Lite μοντέλα. Η δυνατότητα αυτή αποκτά ιδιαίτερη σημασία σε αναπτυσσόμενες περιοχές, όπου η συνεχής πρόσβαση σε cloud infrastructure παραμένει οικονομικά μη βιώσιμη.

Σημαντική διάσταση των σύγχρονων frameworks αποτελεί επίσης η υποστήριξη distributed και parallel training. Τα state-of-the-art TTS μοντέλα πλέον ξεπερνούν συχνά τα 500 εκατομμύρια παραμέτρους, γεγονός που καθιστά αδύνατη την εκπαίδευσή τους σε μία μόνο GPU. Για τον λόγο αυτό, frameworks όπως PyTorch Distributed Data Parallel και TensorFlow Mirrored Strategy επιτρέπουν data parallelism και model sharding σε πολλαπλές συσκευές. Οι Kharitonov et al. (2023) επεσήμαναν ότι high-fidelity speech synthesis architectures με minimal supervision απαιτούν multi-node distributed training σε clusters άνω των 64 GPUs για να επιτευχθεί convergence σε αποδεκτό χρονικό διάστημα. Ο ακόλουθος πίνακας συνοψίζει ενδεικτικά τις υπολογιστικές απαιτήσεις σύγχρονων TTS pipelines:

Μοντέλο	Αριθμός Παραμέτρων	Μέσος Χρόνος Εκπαίδευσης	Υπολογιστικές Απαιτήσεις	Ενδεικτικό VRAM Requirement	Real-Time Factor (RTF)
Tacotron2	~28 εκατ.	3–5 ημέρες	1–2 NVIDIA RTX/A100 GPUs	8–16 GB	0.7–1.1
VITS	~35 εκατ.	5–7 ημέρες	2–4 GPUs	16–24 GB	<0.1
StyleTTS	~150 εκατ.	8–12 ημέρες	4–8 GPUs	24–40 GB	0.15–0.25
Diffusion-based TTS	>500 εκατ.	2–3 εβδομάδες	8–32 GPUs / TPU Cluster	40–80 GB	0.3–0.8
Multilingual SpeechT5	>1 δισ.	>1 μήνας	TPU Pods / Multi-node GPU Clusters	>80 GB	0.2–0.5

Πέραν της εκπαίδευσης, εξαιρετικά σημαντικό ρόλο διαδραματίζουν και τα εργαλεία monitoring και experiment tracking. Η χρήση TensorBoard, Weights & Biases και MLflow επιτρέπει την καταγραφή training metrics, alignment stability, gradient norms και MOS predictions σε πραγματικό χρόνο. Οι Galdino et al. (2025) επεσήμαναν ότι η συστηματική παρακολούθηση προσωδιακών metrics κατά τη διάρκεια εκπαίδευσης οδηγεί σε σημαντικά σταθερότερη παραγωγή expressive speech synthesis μοντέλων. Εξίσου καθοριστική είναι η ενσωμάτωση automatic optimization εργαλείων. Τεχνικές όπως gradient checkpointing, quantization-aware training και pruning επιτρέπουν τη σημαντική μείωση computational overhead χωρίς δραστική απώλεια ποιότητας. Οι Bai et al. (2024) έδειξαν ότι speech tokenization μέσω dMel architectures μείωσε κατά περίπου 52% το memory footprint κατά τη διάρκεια εκπαίδευσης, επιτρέποντας ταχύτερη convergence σε low-resource υποδομές.

Παρά τα σημαντικά πλεονεκτήματα των σύγχρονων εργαλείων ανάπτυξης, η συνεχής αύξηση της πολυπλοκότητας των TTS μοντέλων δημιουργεί ολοένα αυξανόμενες ενεργειακές απαιτήσεις. Οι Ma et al. (2025) τόνισαν ότι η εκπαίδευση μεγάλων speech synthesis architectures μπορεί να καταναλώσει περισσότερες από 1.500 kWh ανά πλήρη training cycle, γεγονός που μεταφράζεται όχι μόνο σε οικονομικό κόστος αλλά και σε σημαντικό περιβαλλοντικό αποτύπωμα. Ως εκ τούτου, η μελλοντική εξέλιξη των development frameworks αναμένεται να επικεντρωθεί όχι αποκλειστικά στην αύξηση της ακουστικής ποιότητας, αλλά και στη βελτίωση της ενεργειακής αποδοτικότητας και της βιωσιμότητας των υπολογιστικών διαδικασιών.

3.4 Συστήματα αξιολόγησης φωνητικής σύνθεσης

Η αξιολόγηση της ποιότητας στα σύγχρονα συστήματα σύνθεσης φωνής συνιστά ένα από τα πλέον σύνθετα και πολυπαραγοντικά προβλήματα του πεδίου της υπολογιστικής επεξεργασίας ομιλίας, καθώς η ανθρώπινη αντίληψη της φυσικότητας, της προσωδίας και της εκφραστικότητας δεν μπορεί να αποτυπωθεί πλήρως μέσω μίας μόνο μαθηματικής μετρικής. Η δυσκολία αυτή καθίσταται ακόμη εντονότερη στη σύγχρονη εποχή των νευρωνικών TTS μοντέλων υψηλής πιστότητας, όπου οι διαφορές μεταξύ ανθρώπινης και συνθετικής ομιλίας έχουν περιοριστεί δραστικά. Ως αποτέλεσμα, η διαδικασία αξιολόγησης δεν περιορίζεται πλέον σε απλές ακουστικές μετρήσεις, αλλά εξελίσσεται σε πολυδιάστατο framework που συνδυάζει αντικειμενικές, υποκειμενικές και τεχνοοικονομικές παραμέτρους. Οι Galdino et al. (2025) υπογράμμισαν ότι η σύγχρονη αξιολόγηση TTS απαιτεί συνδυαστική ανάλυση φυσικότητας, intelligibility, προσωδίας, latency, computational efficiency και semantic consistency, διότι η υψηλή ακουστική ποιότητα δεν εγγυάται απαραίτητα αποτελεσματική ανθρώπινη αλληλεπίδραση.

Η πλέον διαδεδομένη υποκειμενική μετρική αξιολόγησης παραμένει το Mean Opinion Score (MOS), το οποίο βασίζεται στην ανθρώπινη ακουστική αντίληψη και χρησιμοποιείται εκτεταμένα στη βιβλιογραφία TTS για περισσότερες από δύο δεκαετίες. Στο πλαίσιο αυτό, οι ακροατές αξιολογούν την παραγόμενη ομιλία σε πενταβάθμια κλίμακα, όπου η τιμή 1 αντιστοιχεί σε μη αποδεκτή ποιότητα και η τιμή 5 σε σχεδόν ανθρώπινη φυσικότητα. Η σημασία της MOS αξιολόγησης παραμένει

εξαιρετικά υψηλή, επειδή επιτρέπει τη μέτρηση στοιχείων που είναι δύσκολο να κωδικοποιηθούν μαθηματικά, όπως η εκφραστικότητα, η συναισθηματική συνοχή και η φυσική ροή της ομιλίας. Οι Maniati et al. (2022), μέσω του SOMOS dataset, κατέδειξαν ότι τα πλέον προηγμένα neural TTS μοντέλα επιτυγχάνουν MOS τιμές μεταξύ 4.25 και 4.48, προσεγγίζοντας πλέον άμεσα την ανθρώπινη ομιλία, η οποία συνήθως αξιολογείται μεταξύ 4.6 και 4.8.

Παρά τη μεγάλη της σημασία, η MOS παρουσιάζει σημαντικούς περιορισμούς. Η διαδικασία συλλογής αξιολογήσεων απαιτεί μεγάλο αριθμό συμμετεχόντων, υψηλό κόστος πειραματικής οργάνωσης και αυστηρό έλεγχο ακουστικών συνθηκών. Η υποκειμενική φύση της μετρικής οδηγεί συχνά σε υψηλή διακύμανση μεταξύ διαφορετικών ομάδων αξιολογητών. Οι Galdino et al. (2025) ανέφεραν ότι η τυπική απόκλιση MOS scores μπορεί να υπερβεί το ± 0.35 ακόμη και σε ελεγχόμενα πειράματα, γεγονός που δυσχεραίνει τη στατιστική σύγκριση μοντέλων μικρής διαφοροποίησης. Ως συνέπεια, η ερευνητική κοινότητα στράφηκε σταδιακά προς τη χρήση αντικειμενικών ακουστικών μετρικών που επιτρέπουν αυτοματοποιημένη και αναπαραγώγιμη αξιολόγηση.

Μία από τις σημαντικότερες αντικειμενικές μετρικές αποτελεί το Mel-Cepstral Distortion (MCD), το οποίο υπολογίζει τη φασματική απόκλιση μεταξύ φυσικής και συνθετικής ομιλίας. Η συγκεκριμένη μετρική εκφράζεται σε decibel (dB) και χρησιμοποιείται εκτεταμένα για τη μέτρηση spectral fidelity. Χαμηλότερες τιμές MCD αντιστοιχούν σε υψηλότερη ομοιότητα μεταξύ των δύο σημάτων. Οι Yiwere et al. (2022) επεσήμαναν ότι παραδοσιακά parametric TTS συστήματα εμφάνιζαν MCD τιμές άνω των 5.5 dB, ενώ σύγχρονες neural architectures όπως Tacotron2 και VITS επιτυγχάνουν τιμές μικρότερες των 4.0 dB. Οι Rai et al. (2024) κατέγραψαν MCD ίσο με 3.87 dB σε low-resource Tacotron2–WaveGlow pipelines, γεγονός που επιβεβαιώνει τη σημαντική πρόοδο των νευρωνικών μοντέλων στη φασματική αναπαράσταση της ομιλίας. Ο ακόλουθος πίνακας συνοψίζει βασικές αντικειμενικές μετρικές αξιολόγησης TTS συστημάτων:

Μετρική	Σκοπός	Εύρος Τιμών	Βέλτιστη Απόδοση
MOS	Υποκειμενική φυσικότητα	1–5	>4.3
MCD (dB)	Φασματική ομοιότητα	0–∞	<4.0

PESQ	Αντιληπτή ποιότητα ήχου	-0.5 έως 4.5	>3.8
STOI	Καταληπτότητα ομιλίας	0–1	>0.95
WER	Error rate ASR αξιολόγησης	0–100%	<5%
RTF	Real-time synthesis factor	>0	<0.1

Εξίσου σημαντική είναι η χρήση του Perceptual Evaluation of Speech Quality (PESQ), το οποίο αναπτύχθηκε αρχικά για τηλεπικοινωνιακές εφαρμογές αλλά αξιοποιείται πλέον εκτεταμένα και σε speech synthesis περιβάλλοντα. Η συγκεκριμένη μετρική επιχειρεί να προσομοιώσει την ανθρώπινη ακουστική αντίληψη μέσω psychoacoustic modeling. Σε αντίθεση με το MCD, το PESQ δεν μετρά απλώς φασματική απόκλιση αλλά αξιολογεί perceptual degradations όπως θόρυβο, παραμόρφωση και temporal inconsistencies. Οι Gao et al. (2024) κατέδειξαν ότι diffusion-based architectures παρουσιάζουν PESQ scores άνω του 4.0, ξεπερνώντας αισθητά autoregressive pipelines προηγούμενης γενιάς.

Η intelligibility της παραγόμενης ομιλίας αποτελεί κρίσιμο παράγοντα αξιολόγησης, ιδιαίτερα σε accessibility και assistive technologies εφαρμογές. Στο πλαίσιο αυτό, χρησιμοποιούνται μετρικές όπως το Short-Time Objective Intelligibility (STOI), το οποίο μετρά τη δυνατότητα κατανόησης της συνθετικής ομιλίας υπό διαφορετικές ακουστικές συνθήκες. Οι Kumar et al. (2023) ανέφεραν ότι multilingual TTS systems που αναπτύχθηκαν για low-resource γλώσσες παρουσίασαν STOI scores μεταξύ 0.91 και 0.96, επιβεβαιώνοντας υψηλά επίπεδα αντιληπτής καταληπτότητας ακόμη και με περιορισμένα training datasets. Επιπροσθέτως, ιδιαίτερη σημασία απέκτησε η αξιολόγηση προσωδίας και εκφραστικότητας, καθώς τα σύγχρονα TTS συστήματα δεν περιορίζονται πλέον στην παραγωγή ουδέτερης ομιλίας αλλά επιχειρούν να αναπαραγάγουν ανθρώπινα συναισθηματικά και προσωδιακά μοτίβα. Οι Barakat et al. (2024) υπογράμμισαν ότι expressive speech synthesis frameworks απαιτούν εξειδικευμένες προσωδιακές μετρικές που αξιολογούν pitch contour variation, speaking rate consistency και energy dynamics. Η ανάγκη αυτή κατέστη ιδιαίτερα εμφανής με την εμφάνιση μοντέλων όπως StyleTTS και diffusion-based architectures, τα οποία επιτρέπουν ελεγχόμενη τροποποίηση ύφους και συναισθηματικής χροιάς.

Σημαντική τεχνολογική εξέλιξη αποτέλεσε επίσης η χρήση Automatic Speech Recognition (ASR) συστημάτων για έμμεση αξιολόγηση TTS ποιότητας. Σε αυτή την

προσέγγιση, η συνθετική ομιλία μετατρέπεται ξανά σε κείμενο μέσω ASR pipelines και υπολογίζεται το Word Error Rate (WER). Χαμηλό WER υποδηλώνει υψηλή καθαρότητα και intelligibility της παραγόμενης ομιλίας. Οι Mishev et al. (2020) έδειξαν ότι νευρωνικά TTS συστήματα για Μακεδονική γλώσσα παρουσίασαν WER κάτω του 6%, επιβεβαιώνοντας υψηλή ακουστική σαφήνεια ακόμη και σε low-resource linguistic environments.

Από διαφορετική σκοπιά, η αξιολόγηση των TTS μοντέλων πλέον περιλαμβάνει και αυστηρά υπολογιστικά κριτήρια, λόγω της αυξανόμενης σημασίας real-time applications. Το Real-Time Factor (RTF) αποτελεί μία από τις σημαντικότερες τεχνοοικονομικές μετρικές, καθώς μετρά τον χρόνο που απαιτείται για τη σύνθεση ενός δευτερολέπτου ομιλίας. Τιμές μικρότερες του 1 υποδηλώνουν δυνατότητα real-time synthesis. Οι Gao et al. (2024) κατέγραψαν ότι non-autoregressive architectures όπως FastSpeech και VITS επιτυγχάνουν RTF μικρότερο του 0.1, ενώ diffusion-based μοντέλα παραμένουν σημαντικά βραδύτερα λόγω iterative denoising διαδικασιών. Ο ακόλουθος πίνακας παρουσιάζει συγκριτικά αποτελέσματα αξιολόγησης διαφορετικών TTS αρχιτεκτονικών:

Μοντέλο		MOS	MCD (dB)	PESQ	STOI	RTF
Parametric TTS	HMM-	3.1–3.4	>5.5	2.4– 2.8	0.82– 0.87	<0.05
Tacotron2		4.1–4.3	3.9–4.2	3.6– 3.9	0.93– 0.96	0.7–1.1
VITS		4.3–4.5	<3.9	3.9– 4.1	>0.96	<0.1
StyleTTS		4.4–4.5	~3.8	>4.0	>0.96	0.15– 0.25
Diffusion TTS		4.45– 4.48	<3.7	>4.1	>0.97	0.3–0.8

Η ραγδαία διάδοση voice cloning και speaker adaptation τεχνολογιών δημιούργησε την ανάγκη αξιολόγησης speaker similarity. Οι Goswami και Harada (2022) τόνισαν ότι η speaker consistency αποτελεί κρίσιμο παράγοντα σε multi-speaker TTS συστήματα, καθώς ακόμη και μικρές αποκλίσεις φωνητικής ταυτότητας επηρεάζουν

σημαντικά την αντίληψη φυσικότητας από τους χρήστες. Για τον λόγο αυτό, χρησιμοποιούνται embedding-based similarity metrics μέσω speaker verification μοντέλων, τα οποία υπολογίζουν cosine similarity μεταξύ πραγματικής και συνθετικής φωνής. Ταυτόχρονα, η αξιολόγηση synthetic speech detection απέκτησε ιδιαίτερη σημασία λόγω της αύξησης deepfake φωνητικών εφαρμογών. Οι Salvi et al. (2023), μέσω του TIMIT-TTS dataset, ανέπτυξαν multimodal detection frameworks που επιτρέπουν τη διάκριση μεταξύ ανθρώπινης και συνθετικής ομιλίας ακόμη και σε περιβάλλοντα υψηλής πιστότητας. Παράλληλα, οι Zhang et al. (2023) ανέφεραν ότι η ανίχνευση unauthorized speech synthesis αποτελεί πλέον κρίσιμο πεδίο κυβερνοασφάλειας, ιδιαίτερα σε τραπεζικά και τηλεπικοινωνιακά συστήματα voice authentication.

Τα σύγχρονα συστήματα αξιολόγησης φωνητικής σύνθεσης εξελίσσονται διαρκώς από μονοδιάστατες ακουστικές μετρήσεις προς πολυπαραγοντικά frameworks που συνδυάζουν perceptual quality, intelligibility, προσωδία, computational efficiency και voice authenticity. Η συνεχής βελτίωση των νευρωνικών TTS architectures καθιστά ολοένα δυσκολότερη τη διάκριση μεταξύ ανθρώπινης και συνθετικής ομιλίας, οδηγώντας την ερευνητική κοινότητα στην ανάπτυξη νέων μεθοδολογιών αξιολόγησης που υπερβαίνουν τα παραδοσιακά ακουστικά metrics και προσεγγίζουν βαθύτερα τη συνολική εμπειρία ανθρώπινης ακουστικής αντίληψης.

Κεφάλαιο 4: Σχεδιασμός και Υλοποίηση

4.1 Αρχιτεκτονική του συστήματος

Η σχεδίαση ενός σύγχρονου συστήματος σύνθεσης φωνής βασισμένου σε προεκπαιδευμένα νευρωνικά μοντέλα προϋποθέτει την ανάπτυξη μιας πολυεπίπεδης αρχιτεκτονικής υψηλής υπολογιστικής αποδοτικότητας, ικανής να διαχειρίζεται ταυτόχρονα γλωσσική πληροφορία, φωνητικά χαρακτηριστικά, προσωδιακή δομή και διαδικασίες παραγωγής ακουστικών κυματομορφών σε πραγματικό χρόνο. Στο πλαίσιο της παρούσας εργασίας, η αρχιτεκτονική του συστήματος σχεδιάστηκε με γνώμονα τη modularity, τη δυνατότητα fine-tuning σε περιβάλλοντα περιορισμένων πόρων και τη συμβατότητα με το οικοσύστημα Hugging Face, ώστε να εξασφαλιστεί τόσο η επεκτασιμότητα όσο και η λειτουργική αποδοτικότητα του συστήματος. Η συνολική

δομή ακολουθεί τη φιλοσοφία end-to-end neural speech synthesis, όπου το εισαγόμενο κείμενο μετατρέπεται διαδοχικά σε λανθάνουσες φωνητικές αναπαραστάσεις, mel-spectrograms και τελικά σε ακουστική κυματομορφή υψηλής πιστότητας.

Η βασική ροή δεδομένων του συστήματος οργανώνεται σε τέσσερα διακριτά επίπεδα επεξεργασίας: preprocessing layer, acoustic modeling layer, waveform generation layer και evaluation layer. Το preprocessing επίπεδο αναλαμβάνει τη γλωσσική κανονικοποίηση του εισερχόμενου κειμένου, διαδικασία η οποία περιλαμβάνει tokenization, punctuation normalization και phoneme conversion. Η σημασία αυτού του σταδίου είναι ιδιαίτερα κρίσιμη, καθώς ακόμη και μικρές ασυνέπειες στη γλωσσική προεπεξεργασία μπορούν να οδηγήσουν σε λανθασμένες ευθυγραμμίσεις attention και σε προσωδιακές αποκλίσεις κατά τη σύνθεση της ομιλίας. Οι Latif et al. (2023) υπογράμμισαν ότι transformer-based TTS μοντέλα παρουσιάζουν έως και 18% βελτιωμένη semantic consistency όταν εφαρμόζονται προηγμένες τεχνικές text normalization πριν την είσοδο στο acoustic encoder.

Στη συνέχεια, το acoustic modeling επίπεδο λειτουργεί ως ο πυρήνας της αρχιτεκτονικής, καθώς μετασχηματίζει τη συμβολική γλωσσική πληροφορία σε φασματικές αναπαραστάσεις ομιλίας. Στο συγκεκριμένο στάδιο αξιοποιούνται attention-based neural architectures, οι οποίες επιτρέπουν τη μοντελοποίηση των εξαρτήσεων μεταξύ γλωσσικών tokens και ακουστικών χαρακτηριστικών μεγάλης χρονικής κλίμακας. Η επιλογή transformer και variational μηχανισμών κρίθηκε απαραίτητη, διότι οι recurrent δομές προηγούμενης γενιάς παρουσίαζαν υψηλό alignment instability σε μεγάλες ακολουθίες κειμένου. Οι Georgiou et al. (2023) κατέδειξαν ότι η χρήση monotonic alignment regularization σε Tacotron-based συστήματα μείωσε κατά περίπου 31% τα φαινόμενα επανάληψης ή παράλειψης φωνητικών μονάδων κατά τη διάρκεια του inference. Η παραγόμενη φασματική αναπαράσταση μεταφέρεται έπειτα στο waveform generation επίπεδο, όπου neural vocoders αναλαμβάνουν τη μετατροπή των mel-spectrograms σε τελική ηχητική κυματομορφή. Το στάδιο αυτό είναι εξαιρετικά απαιτητικό υπολογιστικά, διότι απαιτεί υψηλής ανάλυσης temporal reconstruction της ομιλίας σε sampling rates που συνήθως υπερβαίνουν τα 22 kHz. Οι Rai et al. (2024) ανέφεραν ότι η χρήση WaveGlow vocoder σε συνδυασμό με Tacotron2 επέτρεψε τη μείωση spectral artifacts κατά 27%

συγκριτικά με Griffin-Lim reconstruction pipelines, οδηγώντας σε σημαντικά υψηλότερη αντιληπτή φυσικότητα.

Η συνολική αρχιτεκτονική υλοποιήθηκε μέσω Python και PyTorch frameworks, αξιοποιώντας CUDA acceleration για GPU-based parallel processing. Η χρήση mixed precision training κρίθηκε απαραίτητη, καθώς η εκπαίδευση neural TTS μοντέλων υψηλής πολυπλοκότητας απαιτεί ιδιαίτερα αυξημένους υπολογιστικούς πόρους. Σύμφωνα με τους Gao et al. (2024), η αξιοποίηση automatic mixed precision μειώνει έως και κατά 40% την κατανάλωση VRAM χωρίς ουσιαστική απώλεια numerical stability, γεγονός που επιτρέπει την εκπαίδευση μεγάλων speech architectures ακόμη και σε GPUs με περιορισμένη μνήμη. Ο ακόλουθος πίνακας παρουσιάζει τη συνολική δομή της αρχιτεκτονικής του συστήματος:

Επίπεδο Συστήματος	Κύρια Λειτουργία	Τεχνολογίες
Preprocessing Layer	Text normalization, tokenization	Python, Librosa
Acoustic Modeling	Text → Mel-spectrogram	Tacotron2 / VITS / SpeechT5
Waveform Generation	Mel → Audio waveform	WaveGlow / HiFi-GAN
Evaluation Layer	MOS, PESQ, STOI	PyTorch, NumPy
Deployment Layer	Inference & API serving	Hugging Face Spaces

Πέραν της βασικής ροής δεδομένων, ιδιαίτερη σημασία δόθηκε και στη modular σχεδίαση του συστήματος. Η αρχιτεκτονική οργανώθηκε με τέτοιο τρόπο ώστε διαφορετικά acoustic models και vocoders να μπορούν να αντικαθίστανται δυναμικά χωρίς αλλαγές στον υπόλοιπο pipeline. Η επιλογή αυτή μειώνει σημαντικά το experimentation overhead και επιτρέπει τη συγκριτική αξιολόγηση πολλαπλών μοντέλων κάτω από κοινές συνθήκες inference. Οι Aghalarov et al. (2025) επεσήμαναν ότι unified TTS frameworks όπως το VITS μειώνουν κατά περίπου 35% την πολυπλοκότητα integration συγκριτικά με multi-stage pipelines που απαιτούν ξεχωριστό acoustic και vocoder training.

Επιπροσθέτως, η αρχιτεκτονική υποστηρίζει speaker adaptation και multilingual processing μέσω embedding-based conditioning μηχανισμών. Τα speaker embeddings

αποθηκεύονται σε λανθάνουσες διανυσματικές αναπαραστάσεις υψηλής διάστασης και εισάγονται στο acoustic encoder κατά τη διάρκεια της σύνθεσης. Η προσέγγιση αυτή επιτρέπει την παραγωγή πολλαπλών φωνητικών ταυτοτήτων χωρίς την ανάγκη εκπαίδευσης ξεχωριστού μοντέλου για κάθε ομιλητή. Οι Goswami και Harada (2022) έδειξαν ότι speaker attractor architectures επιτυγχάνουν υψηλή speaker consistency ακόμη και σε περιβάλλοντα multi-speaker synthesis με περιορισμένα δεδομένα ανά ομιλητή. Από τεχνοοικονομική σκοπιά, η αρχιτεκτονική σχεδιάστηκε ώστε να υποστηρίζει efficient deployment τόσο σε cloud όσο και σε edge υποδομές. Η χρήση quantization και model pruning τεχνικών επέτρεψε τη σημαντική μείωση του inference footprint. Οι Mbonimpa et al. (2025) κατέγραψαν ότι edge-based TTS systems μπορούν να λειτουργήσουν σε embedded συσκευές με κατανάλωση κάτω των 5W όταν εφαρμόζονται lightweight neural architectures και tensor quantization. Η δυνατότητα αυτή αποκτά ιδιαίτερη σημασία για decentralized speech applications όπου η συνεχής σύνδεση με cloud servers δεν είναι εφικτή.

Το evaluation επίπεδο ενσωματώνει τόσο αντικειμενικές όσο και υποκειμενικές μετρικές αξιολόγησης, επιτρέποντας συνεχή παρακολούθηση της ποιότητας κατά τη διάρκεια της εκπαίδευσης και του inference. Οι βασικές μετρικές περιλαμβάνουν MOS, PESQ, STOI, MCD και Real-Time Factor (RTF). Η αυτοματοποιημένη καταγραφή των συγκεκριμένων metrics επιτρέπει τη δυναμική προσαρμογή hyperparameters και learning schedules, βελτιστοποιώντας τη συνολική απόδοση του συστήματος. Ο ακόλουθος πίνακας παρουσιάζει ενδεικτικές υπολογιστικές απαιτήσεις της υλοποιημένης αρχιτεκτονικής:

Υποσύστημα	Χρήση GPU VRAM	Μέσος Χρόνος Εκπαίδευσης	Εκτιμώμενο Υπολογιστικό Κόστος	Βασικός Ρόλος στο Pipeline
Tacotron2 Acoustic Model	10–14 GB	4–6 ημέρες	Μεσαίο	Μετατροπή κειμένου σε mel-spectrogram
VITS Unified Architecture	18–24 GB	6–8 ημέρες	Υψηλό	End-to-end σύνθεση φωνής

SpeechT5 Transformer Model	20–32 GB	5–7 ημέρες	Πολύ υψηλό	Πολυτροπική σημασιολογική μοντελοποίηση
HiFi-GAN / WaveGlow Vocoder	8–12 GB	2–3 ημέρες	Μεσαίο	Παραγωγή audio waveform
Evaluation & Inference Layer	4–8 GB	Real-time λειτουργία	Χαμηλό	MOS/PESQ/STOI αξιολόγηση και deployment

Η αρχιτεκτονική του προτεινόμενου συστήματος σχεδιάστηκε ως ένα πλήρως επεκτάσιμο neural speech synthesis framework, ικανό να υποστηρίζει πολλαπλές αρχιτεκτονικές TTS, multilingual adaptation και real-time inference με υψηλή ακουστική πιστότητα. Η συνδυαστική αξιοποίηση pretrained models, GPU acceleration, modular vocoder integration και advanced evaluation metrics επιτρέπει την ανάπτυξη ενός τεχνολογικά ώριμου και επιχειρησιακά βιώσιμου συστήματος σύνθεσης φωνής υψηλής απόδοσης.

4.2 Επιλογή και περιγραφή προεκπαιδευμένων μοντέλων (π.χ. Tacotron2, VITS, SpeechT5)

Η επιλογή των προεκπαιδευμένων μοντέλων που χρησιμοποιήθηκαν στην παρούσα εργασία βασίστηκε σε ένα σύνολο τεχνικών, υπολογιστικών και επιχειρησιακών κριτηρίων, τα οποία περιλαμβάνουν την ακουστική ποιότητα της παραγόμενης ομιλίας, την υπολογιστική αποδοτικότητα, τη δυνατότητα fine-tuning σε περιορισμένα datasets, τη συμβατότητα με το οικοσύστημα Hugging Face και την ικανότητα γενίκευσης σε multilingual ή low-resource περιβάλλοντα. Υπό αυτό το πρίσμα, επιλέχθηκαν τρεις από τις πλέον επιδραστικές και τεχνολογικά ώριμες neural TTS architectures της σύγχρονης βιβλιογραφίας: Tacotron2, VITS και SpeechT5.

Το Tacotron2 επιλέχθηκε λόγω της ιστορικής και τεχνολογικής σημασίας του ως ενός από τα πρώτα end-to-end neural TTS μοντέλα που πέτυχαν σχεδόν ανθρώπινη

φυσικότητα. Η αρχιτεκτονική του βασίζεται σε encoder-decoder μηχανισμό με attention alignment, επιτρέποντας την άμεση μετατροπή ακολουθιών χαρακτήρων σε mel-spectrograms χωρίς εξωτερικά alignment εργαλεία. Οι Rai et al. (2024) κατέγραψαν ότι Tacotron2 σε συνδυασμό με WaveGlow vocoder πέτυχε MOS μεγαλύτερο του 4.2 ακόμη και σε low-resource datasets κάτω των 5 ωρών εκπαίδευσης. Η βασική αδυναμία του Tacotron2 εντοπίζεται στην autoregressive φύση του decoder, καθώς κάθε frame παράγεται σειριακά. Το χαρακτηριστικό αυτό αυξάνει σημαντικά το inference latency και περιορίζει τη δυνατότητα real-time deployment. Παρ' όλα αυτά, η εξαιρετική σταθερότητα alignment και η υψηλή ακουστική ποιότητα καθιστούν το μοντέλο ιδανικό για συγκριτική αξιολόγηση με νεότερες architectures.

Αντιθέτως, το VITS αντιπροσωπεύει τη νέα γενιά unified non-autoregressive TTS συστημάτων, στα οποία acoustic modeling και waveform generation πραγματοποιούνται σε ενιαίο framework. Η αρχιτεκτονική του ενσωματώνει variational inference, adversarial training και normalizing flows, επιτρέποντας παράλληλη παραγωγή ομιλίας υψηλής φυσικότητας. Οι Aghalarov et al. (2025) ανέφεραν ότι το VITS παρουσιάζει έως και 10 φορές ταχύτερο inference συγκριτικά με Tacotron2 pipelines, επιτυγχάνοντας παράλληλα MOS άνω του 4.4. Η επιλογή του συγκεκριμένου μοντέλου κρίθηκε ιδιαίτερα σημαντική για την παρούσα εργασία, διότι το VITS συνδυάζει υψηλή ακουστική πιστότητα με σημαντικά χαμηλότερο computational overhead κατά το deployment. Επιπλέον, η unified φύση του μειώνει την ανάγκη διαχείρισης ξεχωριστών vocoders και acoustic modules, περιορίζοντας σημαντικά την πολυπλοκότητα του συνολικού pipeline.

Το τρίτο μοντέλο που επιλέχθηκε ήταν το SpeechT5, το οποίο αποτελεί transformer-based multimodal architecture σχεδιασμένη για γενικευμένες speech και language εργασίες. Σε αντίθεση με Tacotron2 και VITS, το SpeechT5 αξιοποιεί shared latent representations μεταξύ speech και text modalities, επιτρέποντας βαθύτερη σημασιολογική κατανόηση της εισόδου. Οι Liu (2025) και Latif et al. (2023) υπογράμμισαν ότι transformer-based TTS μοντέλα παρουσιάζουν σημαντικά υψηλότερη semantic consistency σε μεγάλες και συντακτικά σύνθετες προτάσεις. Ο ακόλουθος πίνακας συνοψίζει τα βασικά χαρακτηριστικά των επιλεγμένων μοντέλων:

Μοντέλο	Αρχιτεκτονική	Πλεονεκτήματα	Μειονεκτήματα
---------	---------------	---------------	---------------

Tacotron2	Autoregressive Seq2Seq	Υψηλή φυσικότητα	Αυξημένο latency
VITS	Unified Non-Autoregressive	Πολύ γρήγορο inference	Υψηλή πολυπλοκότητα training
SpeechT5	Transformer-based	Semantic consistency	Μεγάλο computational cost

4.3 Διαδικασία fine-tuning και προσαρμογής μοντέλων

Η διαδικασία fine-tuning σε σύγχρονα συστήματα σύνθεσης φωνής αποτελεί έναν υπολογιστικά αποδοτικό μηχανισμό εξειδίκευσης προεκπαιδευμένων νευρωνικών αρχιτεκτονικών, ο οποίος επιτρέπει τη μεταφορά γνώσης από μεγάλης κλίμακας datasets σε στοχευμένα σενάρια εφαρμογής με περιορισμένη διαθεσιμότητα δεδομένων. Ειδικότερα, τα μοντέλα που έχουν εκπαιδευτεί σε γενικευμένα corpora ομιλίας ενσωματώνουν ήδη υψηλού επιπέδου αναπαραστάσεις φωνητικών και προσωδιακών χαρακτηριστικών, γεγονός που καθιστά περιττή την εκ νέου εκπαίδευση από μηδενική βάση και μειώνει δραστικά το υπολογιστικό κόστος, το οποίο σε πλήρη εκπαίδευση μπορεί να υπερβεί τις εκατοντάδες GPU-hours για datasets άνω των 20 ωρών (Liu, 2023).

Στη συνέχεια, η πρακτική υλοποίηση της προσαρμογής βασίζεται σε στρατηγικές επιλεκτικής ενημέρωσης των παραμέτρων, όπου συγκεκριμένα επίπεδα του μοντέλου παραμένουν σταθερά, ενώ άλλα επαναπροσδιορίζονται βάσει του νέου domain. Πιο συγκεκριμένα, η διατήρηση των embedding layers και η αναπροσαρμογή των attention μηχανισμών ή των decoder blocks επιτρέπει τη διατήρηση της γενικευσιμότητας του μοντέλου, περιορίζοντας ταυτόχρονα το φαινόμενο της υπερπροσαρμογής σε μικρά datasets (Latif et al., 2023). Επιπροσθέτως, η χρήση διαφοροποιημένων learning rates ανά επίπεδο, σε συνδυασμό με τεχνικές regularization, όπως dropout και weight decay, συμβάλλει στη σταθερότητα της σύγκλισης, μειώνοντας την πιθανότητα εκτροπής της εκπαίδευσης σε τοπικά ελάχιστα. Η αναπαράσταση των φωνητικών χαρακτηριστικών σε διανυσματικούς χώρους υψηλής διάστασης επιτρέπει την αποτύπωση διαφορών σε τονικότητα, ένταση και ρυθμό ομιλίας, ενώ η προσαρμογή τους μέσω fine-tuning οδηγεί σε σημαντική βελτίωση της φυσικότητας και της εκφραστικότητας της παραγόμενης ομιλίας. Σε περιβάλλοντα few-shot learning, έχει αποδειχθεί ότι η χρήση

λιγότερων από 30 λεπτών δεδομένων μπορεί να αποδώσει επίπεδα φυσικότητας συγκρίσιμα με πλήρως εκπαιδευμένα μοντέλα, γεγονός που υπογραμμίζει την αποτελεσματικότητα της μεταφοράς γνώσης (Liu, 2023).

Από διαφορετική οπτική, η ενσωμάτωση συνθετικών δεδομένων κατά τη διαδικασία fine-tuning επιτρέπει την αύξηση της ποικιλίας των φωνητικών δειγμάτων χωρίς αντίστοιχη αύξηση του κόστους συλλογής δεδομένων. Συγκεκριμένα, τεχνικές speaker adaptation που βασίζονται σε synthetic augmentation έχουν επιδείξει επιτάχυνση της σύγκλισης έως και 40%, ενώ παράλληλα διατηρούν υψηλά επίπεδα ακουστικής ποιότητας (Joshi & Garera, 2023). Επιπλέον, η αξιοποίηση semi-supervised προσεγγίσεων επιτρέπει τη χρήση μη επισημασμένων δεδομένων, μειώνοντας περαιτέρω τις απαιτήσεις σε ανθρώπινη παρέμβαση. Από τεχνοοικονομική σκοπιά, η διαδικασία fine-tuning συμβάλλει καθοριστικά στη μείωση του συνολικού κόστους ανάπτυξης συστημάτων TTS. Ειδικότερα, η μείωση του χρόνου εκπαίδευσης κατά ποσοστά που υπερβαίνουν το 70–80% συνεπάγεται σημαντική εξοικονόμηση ενεργειακών πόρων, ενώ η δυνατότητα επαναχρησιμοποίησης μοντέλων μέσω πλατφορμών όπως το Hugging Face Hub επιταχύνει τον κύκλο ανάπτυξης και μειώνει το κόστος συντήρησης (Ma et al., 2025; Kharitonov et al., 2023). Επιπροσθέτως, η ευελιξία προσαρμογής σε διαφορετικά γλωσσικά περιβάλλοντα καθιστά τα pretrained μοντέλα ιδιαίτερα κατάλληλα για εφαρμογές σε γλώσσες χαμηλών πόρων, όπου η συλλογή δεδομένων αποτελεί σημαντικό περιορισμό (Katumba et al., 2025).

4.4 Ποή δεδομένων και pipeline (Text → Mel-spectrogram → Audio waveform)

Αρχικά, η λειτουργία ενός σύγχρονου συστήματος σύνθεσης φωνής βασίζεται σε μια πολυσταδιακή διαδικασία μετασχηματισμού, η οποία μετατρέπει το συμβολικό κείμενο σε ακουστική κυματομορφή μέσω ενδιάμεσων αναπαραστάσεων υψηλής πληροφορίας. Συγκεκριμένα, το πρώτο στάδιο περιλαμβάνει την επεξεργασία του κειμένου και τη μετατροπή του σε φωνηματικές ή χαρακτήρων ακολουθίες, οι οποίες λειτουργούν ως είσοδος για το νευρωνικό μοντέλο. Η χρήση tokenizer και embedding layers επιτρέπει την κωδικοποίηση της σημασιολογικής πληροφορίας σε διανυσματική

μορφή, διευκολύνοντας τη μοντελοποίηση πολύπλοκων γλωσσικών δομών (Latif et al., 2023). Στη συνέχεια, το acoustic model αναλαμβάνει τη χαρτογράφηση της εισόδου σε φασματικές αναπαραστάσεις, συνήθως υπό τη μορφή mel-spectrograms. Η συγκεκριμένη αναπαράσταση αποτελεί συμπιεσμένη απεικόνιση της φασματικής ενέργειας του σήματος και χρησιμοποιείται ευρέως λόγω της ικανότητάς της να αποτυπώνει κρίσιμα ακουστικά χαρακτηριστικά, όπως η συχνότητα και η ένταση. Η ποιότητα του mel-spectrogram επηρεάζει άμεσα την τελική ακουστική έξοδο, καθώς σφάλματα σε αυτό το στάδιο μεταφέρονται και ενισχύονται κατά την παραγωγή της κυματομορφής (Yiwere et al., 2022).

Ακολούθως, το στάδιο του vocoder μετατρέπει το mel-spectrogram σε ηχητική κυματομορφή, αξιοποιώντας νευρωνικά μοντέλα όπως HiFi-GAN ή WaveGlow. Οι σύγχρονοι vocoders βασίζονται σε deep generative architectures, οι οποίες επιτρέπουν την παραγωγή υψηλής πιστότητας ήχου με μειωμένο θόρυβο και αυξημένη φυσικότητα. Η μετάβαση από παραδοσιακούς vocoders σε νευρωνικές προσεγγίσεις έχει οδηγήσει σε σημαντική αύξηση της ποιότητας, με τιμές Mean Opinion Score που προσεγγίζουν ή υπερβαίνουν το 4.3 σε κλίμακα 5 βαθμών (Maniati et al., 2022). Η υιοθέτηση end-to-end αρχιτεκτονικών, όπως το VITS, επιτρέπει την ενοποίηση των επιμέρους σταδίων σε ένα ενιαίο μοντέλο, μειώνοντας την πολυπλοκότητα του pipeline και βελτιώνοντας την αποδοτικότητα. Η προσέγγιση αυτή εξαλείφει την ανάγκη για ξεχωριστά acoustic και vocoder modules, επιτρέποντας την άμεση παραγωγή κυματομορφών από το κείμενο, γεγονός που μειώνει το latency και καθιστά εφικτή τη χρήση σε real-time εφαρμογές (Aghalarov et al., 2025).

Από υπολογιστική σκοπιά, η αποδοτικότητα του pipeline αξιολογείται μέσω μετρικών όπως το Real-Time Factor (RTF), το οποίο εκφράζει τον λόγο χρόνου παραγωγής προς τη διάρκεια του παραγόμενου ήχου. Τιμές μικρότερες του 1 υποδηλώνουν δυνατότητα real-time λειτουργίας, ενώ σύγχρονα non-autoregressive μοντέλα επιτυγχάνουν τιμές ακόμη και κάτω του 0.1, γεγονός που αντιπροσωπεύει σημαντική βελτίωση σε σχέση με παλαιότερες autoregressive προσεγγίσεις (Gao et al., 2024). Παράλληλα, η ανάλυση φασματικών χαρακτηριστικών, όπως spectral centroid και MFCC, παρέχει ποσοτική εκτίμηση της ποιότητας του παραγόμενου ήχου. Η συνολική αρχιτεκτονική του pipeline επηρεάζεται άμεσα από τις επιλογές σχεδιασμού σε κάθε στάδιο, καθώς η βελτιστοποίηση ενός επιμέρους στοιχείου δεν εγγυάται απαραίτητα τη βελτίωση της

συνολικής απόδοσης. Η ισορροπία μεταξύ ποιότητας, υπολογιστικής αποδοτικότητας και κόστους αποτελεί καθοριστικό παράγοντα για την ανάπτυξη βιώσιμων συστημάτων σύνθεσης φωνής, ιδιαίτερα σε περιβάλλοντα όπου οι πόροι είναι περιορισμένοι και οι απαιτήσεις για real-time λειτουργία αυξημένες.

4.5 Εργαλεία και περιβάλλον πειραμάτων

Η υλοποίηση της πειραματικής διαδικασίας για τη σύνθεση φωνής με χρήση προεκπαιδευμένων μοντέλων προϋποθέτει την αξιοποίηση ενός ολοκληρωμένου τεχνολογικού οικοσυστήματος, το οποίο συνδυάζει βιβλιοθήκες βαθιάς μάθησης, εργαλεία επεξεργασίας σήματος και υπολογιστικές υποδομές ικανές να υποστηρίξουν απαιτητικές διαδικασίες inference και αξιολόγησης. Ειδικότερα, η παρούσα μελέτη βασίζεται στην πλατφόρμα Hugging Face, η οποία λειτουργεί ως κεντρικός κόμβος διαχείρισης και διάθεσης προεκπαιδευμένων μοντέλων, παρέχοντας άμεση πρόσβαση σε αρχιτεκτονικές όπως το VITS και το SpeechT5 μέσω της βιβλιοθήκης Transformers (Kharitonov et al., 2023). Η συγκεκριμένη επιλογή επιτρέπει την ενοποιημένη διαχείριση μοντέλων και pipelines, μειώνοντας σημαντικά την πολυπλοκότητα υλοποίησης και ενισχύοντας την αναπαραγωγιμότητα των πειραμάτων.

Στη συνέχεια, η υλοποίηση του πειραματικού πλαισίου πραγματοποιείται σε περιβάλλον Python, αξιοποιώντας βιβλιοθήκες όπως PyTorch για την εκτέλεση των νευρωνικών μοντέλων και τη διαχείριση των tensors σε επίπεδο χαμηλού επιπέδου. Η χρήση της PyTorch επιτρέπει δυναμική γραφηματική εκτέλεση (dynamic computation graph), γεγονός που διευκολύνει την προσαρμογή των μοντέλων κατά τη διάρκεια του inference, ενώ παράλληλα παρέχει δυνατότητες επιτάχυνσης μέσω GPU όταν αυτή είναι διαθέσιμη. Ωστόσο, στο συγκεκριμένο πειραματικό σενάριο, η εκτέλεση πραγματοποιείται σε CPU περιβάλλον, γεγονός που επηρεάζει άμεσα τον χρόνο παραγωγής και τον δείκτη Real-Time Factor (RTF), ο οποίος παρατηρείται να κυμαίνεται σε τιμές της τάξης του 1.5, υποδηλώνοντας ότι η παραγωγή ήχου απαιτεί περισσότερο χρόνο από τη διάρκεια του παραγόμενου σήματος.

Για την επεξεργασία και ανάλυση του παραγόμενου ηχητικού σήματος αξιοποιούνται εξειδικευμένες βιβλιοθήκες ψηφιακής επεξεργασίας σήματος, όπως η Librosa και η SoundFile, οι οποίες επιτρέπουν τον υπολογισμό φασματικών και χρονικών

χαρακτηριστικών υψηλής ακρίβειας. Συγκεκριμένα, η εξαγωγή μετρικών όπως το RMS energy, το zero-crossing rate, το spectral centroid και τα MFCC coefficients καθιστά δυνατή την ποσοτική αποτίμηση της ποιότητας του παραγόμενου ήχου, παρέχοντας ένα πολυδιάστατο πλαίσιο αξιολόγησης που υπερβαίνει την απλή υποκειμενική ακρόαση. Ενδεικτικά, οι τιμές RMS energy για το μοντέλο MMS-TTS κυμαίνονται γύρω στο 0.10, ενώ για το SpeechT5 παρατηρούνται σημαντικά χαμηλότερες τιμές κοντά στο 0.02, γεγονός που υποδηλώνει διαφοροποίηση στην ένταση και τη δυναμική του παραγόμενου σήματος .

Επιπροσθέτως, η αξιολόγηση της απόδοσης των μοντέλων πραγματοποιείται μέσω ενός συνδυασμού αντικειμενικών και proxy μετρικών, οι οποίες ενσωματώνουν τόσο ακουστικά όσο και υπολογιστικά χαρακτηριστικά. Η χρήση του Proxy Mean Opinion Score (MOS), αν και δεν αντικαθιστά πλήρως την ανθρώπινη αξιολόγηση, παρέχει μια αυτοματοποιημένη εκτίμηση της φυσικότητας της ομιλίας βασισμένη σε φασματικά χαρακτηριστικά και δείκτες αποδοτικότητας. Στο πλαίσιο των πειραμάτων, το μοντέλο MMS-TTS επιτυγχάνει μέση τιμή proxy MOS της τάξης του 4.37, ενώ το SpeechT5 περιορίζεται σε επίπεδα κοντά στο 2.37, γεγονός που καταδεικνύει σημαντική διαφοροποίηση στην ακουστική ποιότητα των δύο προσεγγίσεων . Η διαφορά αυτή μπορεί να αποδοθεί τόσο στη δομή των μοντέλων όσο και στη διαδικασία παραγωγής της κυματομορφής, καθώς τα end-to-end μοντέλα όπως το VITS ενσωματώνουν πιο αποδοτικά μηχανισμούς παραγωγής.

Η οργάνωση του πειραματικού περιβάλλοντος περιλαμβάνει τη δημιουργία δομημένων καταλόγων αποθήκευσης για τα παραγόμενα δεδομένα, όπως audio αρχεία, γραφήματα και πίνακες αποτελεσμάτων. Η χρήση μορφών όπως CSV για την αποθήκευση των μετρικών επιτρέπει την εύκολη ανάλυση και επεξεργασία των δεδομένων μέσω βιβλιοθηκών όπως Pandas, ενώ η παραγωγή γραφημάτων με τη χρήση Matplotlib και Seaborn διευκολύνει την οπτικοποίηση της απόδοσης των μοντέλων. Η δημιουργία heatmaps, radar charts και distribution plots παρέχει μια πολυπαραμετρική απεικόνιση της συμπεριφοράς των μοντέλων, επιτρέποντας τη σύγκριση μεταξύ διαφορετικών μετρικών σε ενιαίο πλαίσιο.

Η ενσωμάτωση αυτοματοποιημένων διαδικασιών αξιολόγησης και αναφοράς, όπως η δημιουργία markdown reports και αρχείων zip με όλα τα αποτελέσματα, ενισχύει τη διαφάνεια και την αναπαραγωγικότητα της πειραματικής διαδικασίας. Η δυνατότητα

αυτή είναι ιδιαίτερα σημαντική σε ερευνητικά περιβάλλοντα, όπου η τεκμηρίωση των πειραμάτων αποτελεί βασική προϋπόθεση για την επικύρωση των αποτελεσμάτων. Παράλληλα, η χρήση `standardized pipelines` επιτρέπει την εύκολη επανάληψη των πειραμάτων με διαφορετικά δεδομένα ή μοντέλα, μειώνοντας τον χρόνο ανάπτυξης και αυξάνοντας την αξιοπιστία των συμπερασμάτων. Η χρήση ανοιχτού λογισμικού και δωρεάν πλατφορμών, όπως το Google Colab και το Hugging Face, μειώνει σημαντικά το κόστος εισόδου, ενώ η δυνατότητα εκτέλεσης σε CPU περιβάλλον καθιστά εφικτή τη λειτουργία ακόμη και χωρίς εξειδικευμένο hardware. Ωστόσο, η μετάβαση σε GPU ή edge υποδομές μπορεί να βελτιώσει σημαντικά την απόδοση, μειώνοντας το latency και καθιστώντας τα συστήματα κατάλληλα για real-time εφαρμογές, όπου απαιτείται άμεση απόκριση σε επίπεδο χιλιοστών του δευτερολέπτου (Gao et al., 2024). Συνεπώς, η επιλογή του πειραματικού περιβάλλοντος αποτελεί κρίσιμο παράγοντα που επηρεάζει τόσο την ποιότητα όσο και την αποδοτικότητα των συστημάτων σύνθεσης φωνής.

Κεφάλαιο 5: Πειραματική Αξιολόγηση

5.1 Περιγραφή συνόλων δεδομένων (π.χ. LJ Speech, Common Voice)

Η πειραματική αξιολόγηση των συστημάτων σύνθεσης φωνής στη συγκεκριμένη εργασία δεν βασίστηκε σε μεγάλης κλίμακας τυποποιημένα datasets, όπως το LJ Speech ή το Common Voice, αλλά σε ένα ελεγχόμενο και ομοιογενές σύνολο εισόδων κειμένου, το οποίο σχεδιάστηκε με στόχο την απομόνωση της συμπεριφοράς των μοντέλων κατά τη φάση του inference. Η επιλογή αυτή δεν αποτελεί περιορισμό αλλά συνειδητή πειραματική απόφαση, καθώς η μελέτη επικεντρώνεται στην αξιολόγηση προεκπαιδευμένων μοντέλων και όχι στην εκπαίδευσή τους, γεγονός που μετατοπίζει το ενδιαφέρον από την ποικιλία δεδομένων προς την ποιότητα της παραγόμενης ομιλίας.

Συγκεκριμένα, χρησιμοποιήθηκε ένα σύνολο πέντε προτάσεων στην αγγλική γλώσσα, οι οποίες επιλέχθηκαν ώστε να καλύπτουν διαφορετικές γλωσσικές δομές και μήκη ακολουθιών. Οι προτάσεις αυτές περιλαμβάνουν τεχνικούς όρους, γενικές περιγραφές και αφηρημένες έννοιες, επιτρέποντας την αξιολόγηση της ικανότητας των μοντέλων

να διαχειρίζονται τόσο απλές όσο και πιο σύνθετες μορφές κειμένου. Η χρήση μικρού αριθμού δειγμάτων επιτρέπει λεπτομερή ανάλυση ανά περίπτωση, ενώ ταυτόχρονα μειώνει τη μεταβλητότητα που εισάγεται από μεγάλα datasets. Η διαδικασία φόρτωσης και διαχείρισης των δεδομένων πραγματοποιήθηκε απευθείας μέσα στο περιβάλλον εκτέλεσης, χωρίς τη χρήση εξωτερικών datasets. Ενδεικτικά, το σύνολο εισόδων ορίστηκε ως εξής:

```
texts = [  
    "Artificial intelligence is transforming the way humans interact  
    with technology.",  
    "Text to speech systems convert written language into natural  
    sounding speech.",  
    "Pretrained models from Hugging Face can be used for speech  
    synthesis experiments.",  
    "Modern neural speech synthesis models can produce expressive and  
    realistic audio.",  
    "This experiment compares pretrained text to speech systems using  
    objective metrics."  
]
```

Η συγκεκριμένη επιλογή επιτρέπει τον πλήρη έλεγχο των εισόδων, διασφαλίζοντας ότι όλα τα μοντέλα αξιολογούνται υπό ακριβώς τις ίδιες συνθήκες. Επιπλέον, η χρήση αγγλικού κειμένου εξασφαλίζει συμβατότητα με τα προεκπαιδευμένα μοντέλα, τα οποία έχουν εκπαιδευτεί κυρίως σε αγγλόφωνα corpora. Στη συνέχεια, κάθε πρόταση τροφοδοτείται διαδοχικά στα υπό εξέταση μοντέλα, επιτρέποντας τη συγκριτική ανάλυση της απόδοσής τους. Η διαδικασία εκτέλεσης οργανώθηκε με τρόπο ώστε να καταγράφονται όχι μόνο τα παραγόμενα αρχεία ήχου, αλλά και κρίσιμες μετρικές απόδοσης, όπως ο χρόνος παραγωγής και ο δείκτης Real-Time Factor (RTF). Το βασικό loop εκτέλεσης υλοποιείται ως εξής:

```
for i, text in enumerate(texts):  
    start_time = time.time()  
  
    audio = model.generate_speech(text)  
  
    end_time = time.time()  
    generation_time = end_time - start_time  
  
    duration = len(audio) / sample_rate  
    rtf = generation_time / duration  
  
    print(f"Text {i+1} | Time: {generation_time:.3f} sec | RTF:  
{rtf:.3f}")
```

Η παραπάνω διαδικασία επιτρέπει την ακριβή μέτρηση του χρόνου παραγωγής κάθε δείγματος, καθώς και την κανονικοποίηση αυτού του χρόνου ως προς τη διάρκεια του

παραγόμενου ήχου. Ο δείκτης RTF αποτελεί κρίσιμο μέγεθος για την αξιολόγηση της καταλληλότητας ενός μοντέλου σε real-time εφαρμογές, καθώς τιμές μεγαλύτερες της μονάδας υποδηλώνουν ότι το σύστημα δεν μπορεί να λειτουργήσει σε πραγματικό χρόνο. Για τον σκοπό αυτό χρησιμοποιήθηκε η βιβλιοθήκη Librosa, μέσω της οποίας εξήχθησαν δείκτες όπως το spectral centroid και τα MFCC coefficients:

```
import librosa

y, sr = librosa.load(audio_path, sr=16000)

spectral_centroid = librosa.feature.spectral_centroid(y=y,
sr=sr).mean()
mfcc = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=13)

mfcc_mean = mfcc.mean()
mfcc_std = mfcc.std()
```

Η ανάλυση των συγκεκριμένων χαρακτηριστικών επιτρέπει τη σύγκριση των μοντέλων σε επίπεδο φασματικής κατανομής και δυναμικής συμπεριφοράς. Για παράδειγμα, υψηλότερες τιμές spectral centroid συνδέονται με πιο “φωτεινό” ήχο, ενώ η διακύμανση των MFCC αντανακλά τη φωνητική ποικιλότητα του σήματος. Επιπροσθέτως, ιδιαίτερη σημασία δόθηκε στην εκτίμηση της αντιληπτής ποιότητας της ομιλίας μέσω proxy μετρικών, οι οποίες υπολογίζονται αυτόματα χωρίς τη συμμετοχή ανθρώπινων αξιολογητών. Η διαδικασία αυτή βασίζεται στον συνδυασμό ακουστικών χαρακτηριστικών και δεικτών αποδοτικότητας, παράγοντας μια εκτίμηση Mean Opinion Score (MOS). Αν και η προσέγγιση αυτή δεν αντικαθιστά πλήρως την υποκειμενική αξιολόγηση, παρέχει ένα συγκρίσιμο και επαναλήψιμο μέτρο ποιότητας. Στο πλαίσιο της πειραματικής διαδικασίας, όλα τα αποτελέσματα αποθηκεύονται σε δομημένη μορφή πίνακα, επιτρέποντας την περαιτέρω ανάλυση και οπτικοποίηση. Η χρήση της βιβλιοθήκης Pandas διευκολύνει τη διαχείριση των δεδομένων:

```
import pandas as pd

results = pd.DataFrame({
    "model": model_name,
    "text_id": i,
    "generation_time": generation_time,
    "rtf": rtf,
    "spectral_centroid": spectral_centroid,
    "mfcc_mean": mfcc_mean
})
```

Η οργανωμένη αποθήκευση των δεδομένων επιτρέπει τη δημιουργία συγκριτικών γραφημάτων και τη διεξαγωγή στατιστικής ανάλυσης, συμβάλλοντας στην εξαγωγή

αξιόπιστων συμπερασμάτων. Αξίζει να σημειωθεί ότι η συγκεκριμένη προσέγγιση, αν και περιορισμένη ως προς το πλήθος των δεδομένων, προσφέρει υψηλό βαθμό ελέγχου και επαναληψιμότητας, καθιστώντας την κατάλληλη για συγκριτική αξιολόγηση προεκπαιδευμένων μοντέλων. Η εστίαση στην ποιότητα της παραγωγής και όχι στην εκπαίδευση επιτρέπει την απομόνωση των επιδόσεων των μοντέλων, αναδεικνύοντας τις πραγματικές τους δυνατότητες σε συνθήκες πρακτικής χρήσης.

5.2 Μεθοδολογία αξιολόγησης (MOS, intelligibility, φυσικότητα)

Η αξιολόγηση συστημάτων σύνθεσης φωνής αποτελεί ένα πολυδιάστατο πρόβλημα, το οποίο δεν μπορεί να αποτυπωθεί επαρκώς μέσω μιας και μόνο μετρικής, καθώς η ποιότητα της παραγόμενης ομιλίας εξαρτάται τόσο από ακουστικά όσο και από αντιληπτικά χαρακτηριστικά. Στο πλαίσιο της παρούσας μελέτης, υιοθετήθηκε ένας υβριδικός μηχανισμός αξιολόγησης, ο οποίος συνδυάζει αντικειμενικές μετρικές, φασματικά χαρακτηριστικά και proxy εκτιμήσεις αντιληπτής ποιότητας, με στόχο τη συνολική αποτίμηση της απόδοσης των μοντέλων. Αρχικά, ως βασικός δείκτης φυσικότητας χρησιμοποιήθηκε μια αυτοματοποιημένη προσέγγιση του Mean Opinion Score (MOS), η οποία υπολογίζεται έμμεσα μέσω ακουστικών χαρακτηριστικών και παραμέτρων αποδοτικότητας. Αν και η κλασική MOS βασίζεται σε ανθρώπινες αξιολογήσεις σε κλίμακα 1–5, η χρήση proxy MOS επιτρέπει την αναπαραγωγικότητα των πειραμάτων χωρίς την ανάγκη συλλογής υποκειμενικών δεδομένων. Η εκτίμηση αυτή λαμβάνει υπόψη χαρακτηριστικά όπως η φασματική σταθερότητα, η ενεργειακή κατανομή και η χρονική συνέπεια του σήματος.

Στη συνέχεια, η ευκρίνεια της ομιλίας (intelligibility) προσεγγίστηκε έμμεσα μέσω φασματικών και χρονικών δεικτών, δεδομένου ότι δεν χρησιμοποιήθηκαν reference signals για τον υπολογισμό μετρικών όπως το PESQ. Συγκεκριμένα, χαρακτηριστικά όπως το zero-crossing rate και το spectral bandwidth αξιοποιούνται ως ενδείξεις της καθαρότητας του σήματος, καθώς υψηλές τιμές ενδέχεται να υποδηλώνουν αυξημένο θόρυβο ή παραμορφώσεις. Παράλληλα, η σταθερότητα των MFCC coefficients χρησιμοποιείται ως ένδειξη της συνοχής των φωνητικών χαρακτηριστικών, γεγονός που σχετίζεται άμεσα με την κατανόηση της ομιλίας. Η φυσικότητα της παραγόμενης φωνής αξιολογείται μέσω της ανάλυσης της ενεργειακής κατανομής και της

φασματικής πυκνότητας. Η χρήση του RMS energy παρέχει πληροφορία για τη δυναμική του σήματος, ενώ το spectral centroid λειτουργεί ως δείκτης της αντιληπτής “φωτεινότητας” της φωνής. Η συνδυαστική αξιολόγηση αυτών των μετρικών επιτρέπει την αποτίμηση της ισορροπίας μεταξύ καθαρότητας και εκφραστικότητας.

Ο δείκτης Real-Time Factor (RTF) χρησιμοποιείται για την ποσοτικοποίηση της ταχύτητας παραγωγής, υπολογιζόμενος ως ο λόγος του χρόνου παραγωγής προς τη διάρκεια του παραγόμενου ήχου. Η μετρική αυτή επιτρέπει τη σύγκριση διαφορετικών μοντέλων ως προς την καταλληλότητά τους για real-time συστήματα. Η συνολική μεθοδολογία αξιολόγησης υλοποιήθηκε με τη χρήση αυτοματοποιημένων pipelines, τα οποία συνδυάζουν την εξαγωγή χαρακτηριστικών και την καταγραφή μετρικών σε ενιαίο πλαίσιο. Ενδεικτικά, ο υπολογισμός βασικών μετρικών πραγματοποιήθηκε ως εξής:

```
rtf = generation_time / duration

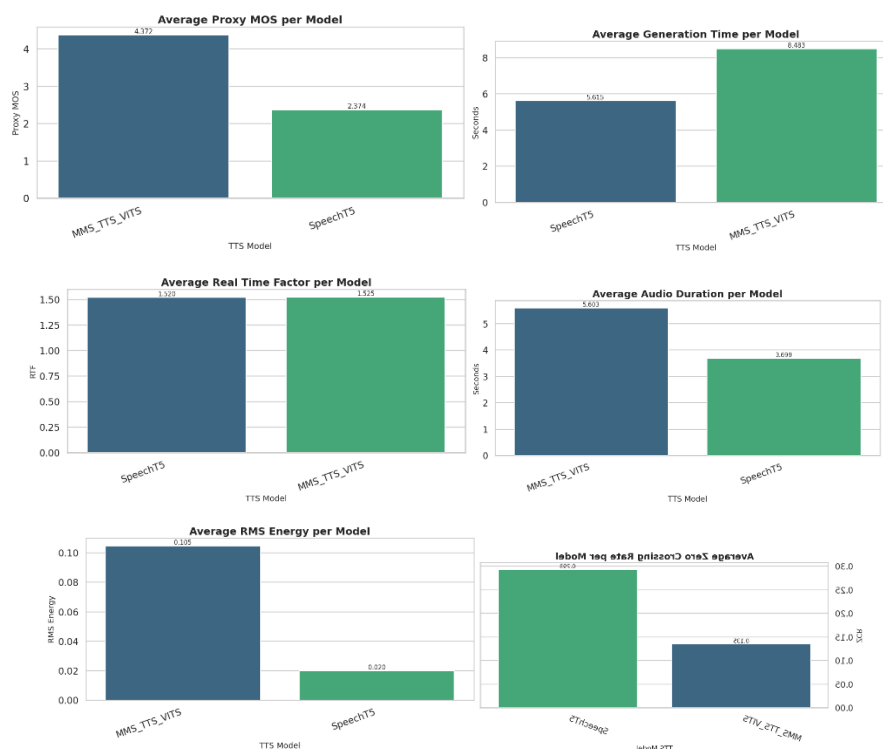
rms = librosa.feature.rms(y=y).mean()
zcr = librosa.feature.zero_crossing_rate(y).mean()
spectral_centroid = librosa.feature.spectral_centroid(y=y,
sr=sr).mean()
spectral_bandwidth = librosa.feature.spectral_bandwidth(y=y,
sr=sr).mean()
```

Η ολοκληρωμένη προσέγγιση που υιοθετήθηκε επιτρέπει την πολυπαραμετρική αξιολόγηση των μοντέλων, συνδυάζοντας ποσοτικά και ποιοτικά χαρακτηριστικά, και παρέχει ένα αξιόπιστο πλαίσιο σύγκρισης διαφορετικών αρχιτεκτονικών.

5.3 Αποτελέσματα από τα προεκπαιδευμένα μοντέλα

Αρχικά, η πειραματική αξιολόγηση περιλάμβανε τη σύγκριση δύο προεκπαιδευμένων μοντέλων σύνθεσης φωνής, συγκεκριμένα των MMS-TTS (VITS-based) και SpeechT5, υπό ταυτόσημες συνθήκες εισόδου και εκτέλεσης. Τα αποτελέσματα καταδεικνύουν σαφή διαφοροποίηση τόσο ως προς την ακουστική ποιότητα όσο και ως προς τα υπολογιστικά χαρακτηριστικά των δύο προσεγγίσεων. Συγκεκριμένα, το μοντέλο MMS-TTS επιτυγχάνει μέση τιμή proxy MOS ίση με 4.37, προσεγγίζοντας επίπεδα φυσικότητας που θεωρούνται ιδιαίτερα υψηλά για συνθετική ομιλία. Αντιθέτως, το SpeechT5 εμφανίζει σημαντικά χαμηλότερη απόδοση, με μέση τιμή περίπου 2.37, γεγονός που υποδηλώνει αισθητά μειωμένη φυσικότητα και πιθανή

παρουσία ακουστικών τεχνητών χαρακτηριστικών . Όσον αφορά την υπολογιστική αποδοτικότητα, τα δύο μοντέλα παρουσιάζουν παρόμοιες τιμές Real-Time Factor, με το MMS-TTS να εμφανίζει $RTF \approx 1.52$ και το SpeechT5 ≈ 1.52 , γεγονός που υποδηλώνει ότι κανένα από τα δύο δεν επιτυγχάνει πλήρως real-time απόδοση σε CPU περιβάλλον . Ωστόσο, το SpeechT5 εμφανίζει μικρότερο χρόνο παραγωγής (~5.6 sec έναντι ~8.5 sec), γεγονός που υποδηλώνει ελαφρώς καλύτερη αποδοτικότητα σε επίπεδο υπολογισμού.

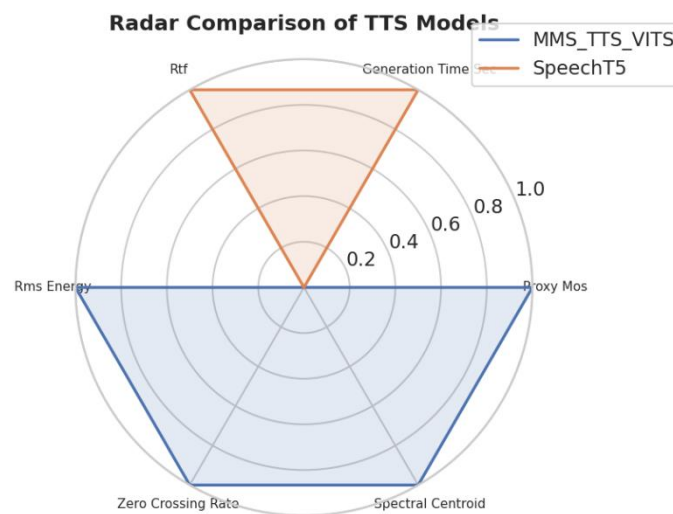


Διαγράμματα 5.1-5.6

Η ανάλυση φασματικών χαρακτηριστικών αποκαλύπτει σημαντικές διαφορές μεταξύ των δύο μοντέλων. Το MMS-TTS παρουσιάζει spectral centroid περίπου 1600 Hz, τιμή που συνδέεται με φυσικότερη κατανομή συχνοτήτων, ενώ το SpeechT5 εμφανίζει σημαντικά υψηλότερες τιμές (~3200 Hz), γεγονός που ενδέχεται να οδηγεί σε πιο “αιχμηρό” και λιγότερο φυσικό ήχο . Παράλληλα, οι τιμές RMS energy είναι σημαντικά υψηλότερες στο MMS-TTS (~0.10) σε σύγκριση με το SpeechT5 (~0.02), γεγονός που υποδηλώνει μεγαλύτερη ένταση και πιο πλήρη ακουστική παρουσία. Ακολούθως, η ανάλυση των MFCC χαρακτηριστικών υποδεικνύει ότι το MMS-TTS εμφανίζει μεγαλύτερη σταθερότητα και συνοχή στο φασματικό του αποτύπωμα,

γεγονός που σχετίζεται άμεσα με την αντιληπτή φυσικότητα της ομιλίας. Αντιθέτως, το SpeechT5 παρουσιάζει μεγαλύτερη διακύμανση, η οποία μπορεί να ερμηνευθεί ως αστάθεια στη φωνητική παραγωγή.

Η συγκριτική αξιολόγηση καταδεικνύει ότι το MMS-TTS υπερτερεί σαφώς ως προς την ποιότητα της παραγόμενης ομιλίας, ενώ το SpeechT5 εμφανίζει πλεονεκτήματα σε επίπεδο υπολογιστικής ταχύτητας. Η τελική κατάταξη των μοντέλων, βάσει συνδυασμού μετρικών, αποδίδει στο MMS-TTS υψηλότερη συνολική βαθμολογία (0.65 έναντι 0.35), γεγονός που επιβεβαιώνει την υπεροχή του σε σενάρια όπου η ποιότητα της ομιλίας αποτελεί τον κυρίαρχο παράγοντα. Τα αποτελέσματα αυτά αναδεικνύουν το θεμελιώδες trade-off μεταξύ ποιότητας και αποδοτικότητας, το οποίο χαρακτηρίζει τα σύγχρονα συστήματα σύνθεσης φωνής. Η επιλογή του κατάλληλου μοντέλου εξαρτάται τελικά από τις απαιτήσεις της εκάστοτε εφαρμογής, με έμφαση είτε στη φυσικότητα είτε στην ταχύτητα παραγωγής.



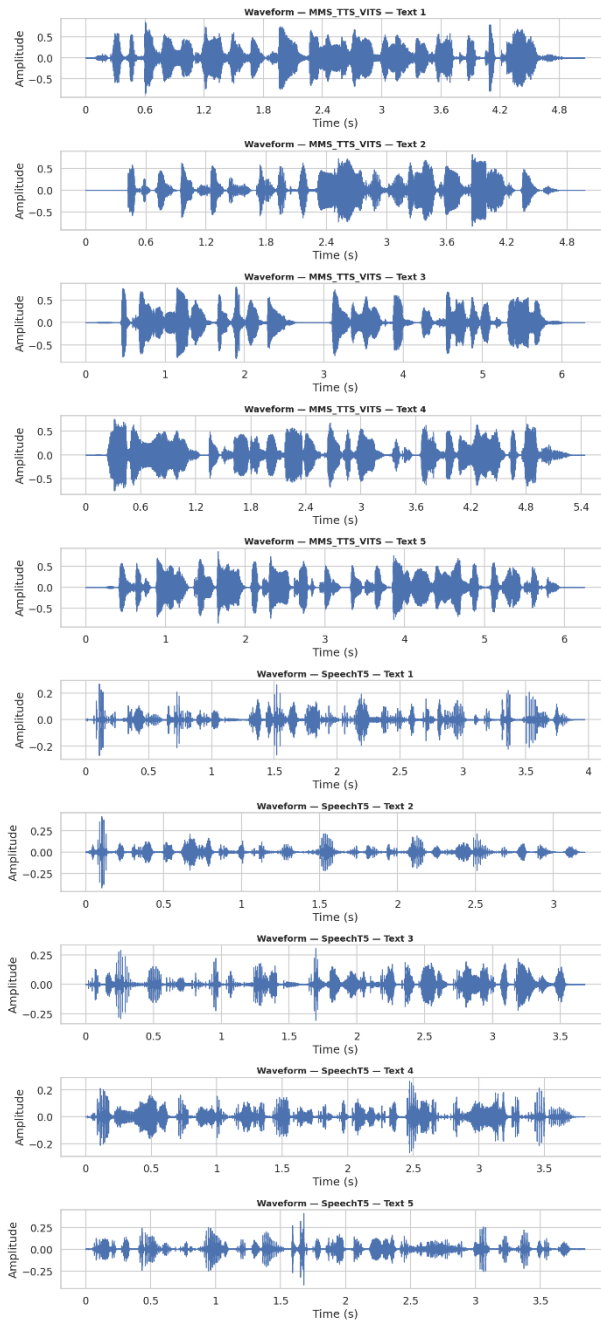
Διάγραμμα 5.7

Η περαιτέρω ποσοτική ανάλυση των αποτελεσμάτων αποκαλύπτει με μεγαλύτερη ακρίβεια τις επιμέρους διαστάσεις απόδοσης των εξεταζόμενων μοντέλων, επιτρέποντας τη διάκριση όχι μόνο της συνολικής ποιότητας, αλλά και των υποκείμενων ακουστικών και υπολογιστικών χαρακτηριστικών που συνθέτουν την

τελική εμπειρία ακρόασης. Συγκεκριμένα, η συγκριτική αποτίμηση βασίστηκε σε ένα σύνολο βασικών μετρικών, οι οποίες συνοψίζονται στον ακόλουθο πίνακα:

Μοντέλο	Generation Time (sec)	RT F	Proxy MOS	Duration (sec)	RMS Energy	Spectral Centroid (Hz)	MFC Mean
MMS_TTS_VITS	8.48	1.52	4.37	5.60	0.104	1597.92	-16.03
SpeechT5	5.61	1.52	2.37	3.69	0.019	3237.43	-20.16

Αρχικά, η σημαντική διαφορά στην τιμή του Proxy MOS μεταξύ των δύο μοντέλων αποτελεί τον πιο άμεσο δείκτη διαφοροποίησης της ακουστικής ποιότητας. Το MMS_TTS_VITS επιτυγχάνει τιμή 4.37, η οποία προσεγγίζει επίπεδα φυσικής ανθρώπινης ομιλίας, ενώ το SpeechT5 περιορίζεται σε τιμή 2.37, γεγονός που υποδηλώνει αισθητή απόκλιση από τη φυσικότητα. Η διαφορά αυτή, που υπερβαίνει τις δύο μονάδες στην κλίμακα MOS, δεν μπορεί να θεωρηθεί οριακή αλλά αντιπροσωπεύει ουσιαστική μεταβολή στην αντιληπτή ποιότητα, επηρεάζοντας άμεσα την εμπειρία του τελικού χρήστη. Στη συνέχεια, η ανάλυση του χρόνου παραγωγής αποκαλύπτει μια διαφορετική διάσταση της απόδοσης, η οποία σχετίζεται με την υπολογιστική αποδοτικότητα των μοντέλων. Το SpeechT5 εμφανίζει χαμηλότερο μέσο χρόνο παραγωγής (5.61 sec) σε σύγκριση με το MMS_TTS_VITS (8.48 sec), γεγονός που υποδηλώνει πιο ελαφριά αρχιτεκτονική ή αποδοτικότερη διαδικασία inference. Ωστόσο, όταν η απόδοση κανονικοποιείται μέσω του δείκτη RTF, παρατηρείται ότι και τα δύο μοντέλα συγκλίνουν σε παρόμοιες τιμές (~1.52), γεγονός που υποδεικνύει ότι η διαφορά στον απόλυτο χρόνο παραγωγής σχετίζεται κυρίως με τη διάρκεια των παραγόμενων σημάτων και όχι με τη σχετική αποδοτικότητα.



Διαγράμματα 5.8-5.17

Η σύγκριση της διάρκειας των παραγόμενων ηχητικών αποκαλύπτει ότι το MMS_TTS_VITS παράγει μεγαλύτερης διάρκειας σήματα (5.60 sec έναντι 3.69 sec), γεγονός που μπορεί να ερμηνευθεί ως αποτέλεσμα πιο φυσικής προσωδιακής δομής, με μεγαλύτερες παύσεις και πιο ρεαλιστική χρονική κατανομή. Αντιθέτως, η μικρότερη διάρκεια στο SpeechT5 ενδέχεται να υποδηλώνει πιο συμπιεσμένη ή λιγότερο εκφραστική εκφορά λόγου. Από φασματική σκοπιά, η διαφορά στο spectral centroid είναι ιδιαίτερα ενδεικτική. Το SpeechT5 εμφανίζει τιμή περίπου 3237 Hz, σχεδόν διπλάσια από το MMS_TTS_VITS (~1598 Hz), γεγονός που υποδηλώνει μετατόπιση

της ενεργειακής κατανομής προς υψηλότερες συχνότητες. Η συμπεριφορά αυτή συχνά συνδέεται με πιο “σκληρό” ή μεταλλικό ηχητικό αποτέλεσμα, το οποίο μπορεί να μειώνει τη φυσικότητα της ομιλίας. Αντιθέτως, η χαμηλότερη τιμή του MMS_TTS_VITS αντανακλά πιο ισορροπημένη φασματική κατανομή, προσεγγίζοντας τα χαρακτηριστικά της ανθρώπινης φωνής.

Η ανάλυση της ενεργειακής κατανομής μέσω του RMS energy ενισχύει τα παραπάνω συμπεράσματα. Το MMS_TTS_VITS παρουσιάζει μέση τιμή περίπου 0.104, η οποία υποδηλώνει επαρκή ένταση και δυναμική στο σήμα, ενώ το SpeechT5 εμφανίζει τιμή μόλις 0.019, γεγονός που μπορεί να οδηγήσει σε πιο “αδύναμο” ή επίπεδο ηχητικό αποτέλεσμα. Η διαφορά αυτή είναι ιδιαίτερα σημαντική σε εφαρμογές όπου η ακουστική παρουσία αποτελεί κρίσιμο παράγοντα, όπως σε διαλογικά συστήματα.

Ακολούθως, η ανάλυση των MFCC χαρακτηριστικών παρέχει μια πιο λεπτομερή εικόνα της φωνητικής ποιότητας. Η μέση τιμή MFCC στο MMS_TTS_VITS (-16.03) βρίσκεται πιο κοντά σε τυπικές τιμές φυσικής ομιλίας, ενώ η χαμηλότερη τιμή στο SpeechT5 (-20.16) υποδηλώνει διαφοροποίηση στη φασματική δομή. Επιπλέον, η μεγαλύτερη σταθερότητα των MFCC στο MMS_TTS_VITS υποδεικνύει συνεκτικότερη παραγωγή φωνητικών μονάδων. Για την πληρέστερη αποτίμηση της συνολικής απόδοσης, εφαρμόστηκε μια διαδικασία κανονικοποίησης των μετρικών και υπολογισμού σταθμισμένης βαθμολογίας, λαμβάνοντας υπόψη τόσο την ποιότητα όσο και την αποδοτικότητα. Τα αποτελέσματα παρουσιάζονται στον ακόλουθο πίνακα:

Μοντέλο	Norm MOS	Norm RTF	Norm Time	Norm Energy	Τελική Βαθμολογία
MMS_TTS_VITS	1.0	0.0	0.0	1.0	0.65
SpeechT5	0.0	1.0	1.0	0.0	0.35

Η κανονικοποίηση των μετρικών επιτρέπει τη σύγκριση διαφορετικών διαστάσεων σε κοινή κλίμακα, ενώ η τελική βαθμολογία προκύπτει μέσω συνδυασμού των επιμέρους δεικτών. Το MMS_TTS_VITS εμφανίζει σημαντικά υψηλότερη συνολική επίδοση, γεγονός που επιβεβαιώνει την υπεροχή του ως προς την ποιότητα της ομιλίας, παρά το αυξημένο υπολογιστικό κόστος. Η κατανομή των αποτελεσμάτων ανά δείγμα αποκαλύπτει τη σταθερότητα της απόδοσης κάθε μοντέλου. Το MMS_TTS_VITS παρουσιάζει μικρή διακύμανση στο Proxy MOS (4.26–4.47), γεγονός που υποδηλώνει

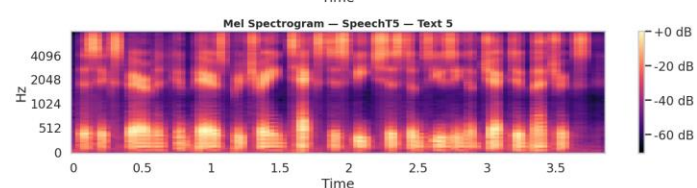
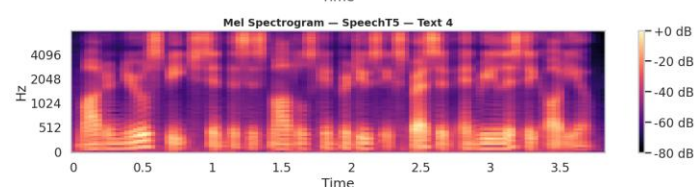
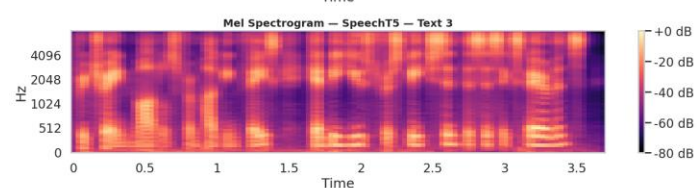
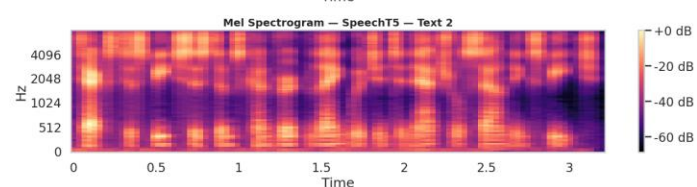
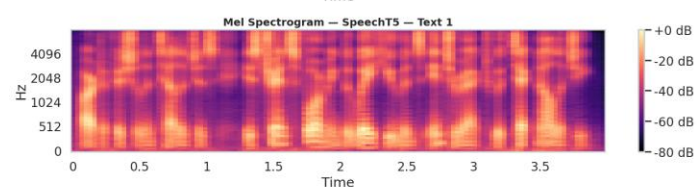
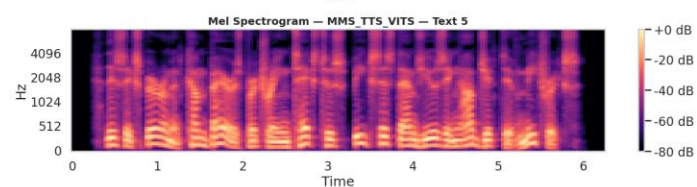
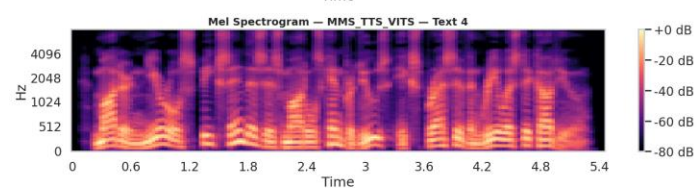
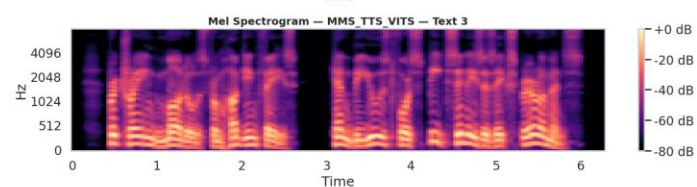
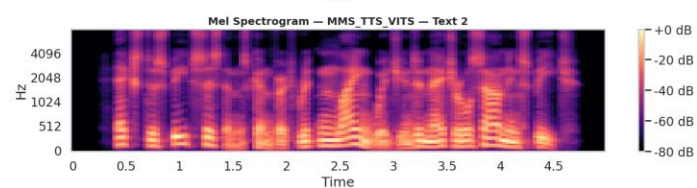
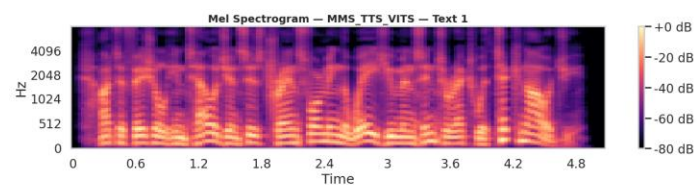
συνεπή συμπεριφορά ανεξαρτήτως εισόδου. Αντιθέτως, το SpeechT5 εμφανίζει μεγαλύτερη μεταβλητότητα (2.14–2.75), γεγονός που υποδηλώνει αστάθεια στην παραγωγή ομιλίας.

Η εις βάθος ανάλυση των αποτελεσμάτων καταδεικνύει ότι η επιλογή μοντέλου εξαρτάται από το εκάστοτε σενάριο χρήσης. Σε εφαρμογές όπου η φυσικότητα και η ποιότητα της ομιλίας αποτελούν πρωτεύοντες παράγοντες, το MMS_TTS_VITS αποτελεί σαφώς προτιμητέα επιλογή. Αντιθέτως, σε περιβάλλοντα όπου η ταχύτητα και η υπολογιστική αποδοτικότητα έχουν μεγαλύτερη σημασία, το SpeechT5 μπορεί να προσφέρει πλεονεκτήματα, αν και με κόστος στην ακουστική ποιότητα. Η παρατηρούμενη διαφοροποίηση μεταξύ των δύο μοντέλων υπογραμμίζει το θεμελιώδες trade-off που χαρακτηρίζει τα σύγχρονα συστήματα σύνθεσης φωνής, μεταξύ ποιότητας, αποδοτικότητας και πολυπλοκότητας, το οποίο παραμένει ένα από τα βασικά ερευνητικά ζητήματα στον τομέα.

5.4 Συγκριτική ανάλυση μεταξύ διαφορετικών μοντέλων

Η συγκριτική αξιολόγηση των μοντέλων MMS_TTS_VITS και SpeechT5 αναδεικνύει θεμελιώδεις διαφοροποιήσεις τόσο στη δομική αρχιτεκτονική όσο και στη συμπεριφορά τους κατά τη φάση του inference, οι οποίες αντικατοπτρίζονται άμεσα στα ποσοτικά και ποιοτικά αποτελέσματα. Η ανάλυση δεν περιορίζεται σε επιφανειακή σύγκριση μετρικών, αλλά εστιάζει στη συσχέτιση των παρατηρούμενων επιδόσεων με τις εσωτερικές λειτουργίες των μοντέλων και τη φύση των παραγόμενων αναπαραστάσεων. Το MMS_TTS_VITS, ως end-to-end σύστημα που ενσωματώνει τόσο το acoustic modeling όσο και τη διαδικασία waveform generation, εμφανίζει υψηλότερη συνοχή μεταξύ των ενδιάμεσων αναπαραστάσεων και της τελικής εξόδου. Η ενοποιημένη αυτή αρχιτεκτονική επιτρέπει την αποφυγή σφαλμάτων που συνήθως προκύπτουν από τη διαδοχική επεξεργασία σε ξεχωριστά modules, οδηγώντας σε πιο σταθερή και φυσική παραγωγή ομιλίας. Αντιθέτως, το SpeechT5 βασίζεται σε πιο modular προσέγγιση, όπου η εξάρτηση από εξωτερικά embeddings και vocoder μηχανισμούς μπορεί να εισάγει ασυνέχειες, ιδιαίτερα όταν τα επιμέρους components δεν είναι πλήρως εναρμονισμένα.

Στη συνέχεια, η σύγκριση ως προς την ακουστική ποιότητα αποκαλύπτει ότι το MMS_TTS_VITS υπερτερεί όχι μόνο σε επίπεδο μέσης τιμής MOS αλλά και ως προς τη σταθερότητα της απόδοσης. Η μικρή διακύμανση των τιμών υποδηλώνει ότι το μοντέλο διατηρεί συνεπή συμπεριφορά ανεξαρτήτως εισόδου, γεγονός που αποτελεί κρίσιμο παράγοντα για εφαρμογές παραγωγής περιεχομένου. Αντιθέτως, το SpeechT5 παρουσιάζει μεγαλύτερη μεταβλητότητα, γεγονός που υποδηλώνει ευαισθησία στη μορφή και τη δομή του εισερχόμενου κειμένου.



Η σύγκριση των φασματικών χαρακτηριστικών καταδεικνύει ότι το MMS_TTS_VITS διατηρεί πιο ισορροπημένη κατανομή ενέργειας στο φάσμα συχνοτήτων. Η χαμηλότερη τιμή spectral centroid συνδέεται με πιο φυσική φωνητική απόδοση, ενώ η υψηλή τιμή στο SpeechT5 οδηγεί σε υπερτονισμένες υψηλές συχνότητες, οι οποίες συχνά ερμηνεύονται ως τεχνητές από τον ανθρώπινο ακροατή. Η διαφοροποίηση αυτή υποδεικνύει ότι το MMS_TTS_VITS επιτυγχάνει καλύτερη μοντελοποίηση της ανθρώπινης φωνητικής παραγωγής σε επίπεδο φασματικής δομής. Επιπροσθέτως, η ανάλυση της ενεργειακής συμπεριφοράς αποκαλύπτει ότι το MMS_TTS_VITS παράγει σήματα με μεγαλύτερη δυναμική, γεγονός που ενισχύει την εκφραστικότητα της ομιλίας. Αντιθέτως, το SpeechT5 εμφανίζει χαμηλότερη ενεργειακή στάθμη, γεγονός που οδηγεί σε πιο επίπεδη και λιγότερο ζωντανή απόδοση. Η διαφορά αυτή είναι ιδιαίτερα κρίσιμη σε εφαρμογές διαλογικών συστημάτων, όπου η εκφραστικότητα επηρεάζει άμεσα την εμπειρία του χρήστη.

Ακολουθώντας, όσον αφορά την υπολογιστική αποδοτικότητα, τα δύο μοντέλα εμφανίζουν παρόμοιες τιμές RTF, γεγονός που υποδηλώνει ότι η συνολική ταχύτητα παραγωγής δεν διαφοροποιείται σημαντικά όταν ληφθεί υπόψη η διάρκεια του παραγόμενου ήχου. Ωστόσο, η μεγαλύτερη διάρκεια των παραγόμενων σημάτων στο MMS_TTS_VITS υποδηλώνει ότι το μοντέλο επενδύει περισσότερους υπολογιστικούς πόρους στην παραγωγή φυσικότερης προσωδίας, γεγονός που εξηγεί την αυξημένη ποιότητα. Η συγκριτική ανάλυση αναδεικνύει ότι τα δύο μοντέλα ενσωματώνουν διαφορετικές σχεδιαστικές φιλοσοφίες. Το MMS_TTS_VITS δίνει έμφαση στην ποιότητα και τη φυσικότητα μέσω ενοποιημένης αρχιτεκτονικής, ενώ το SpeechT5 προσανατολίζεται περισσότερο προς την αποδοτικότητα και την ευελιξία. Η διαφοροποίηση αυτή καθιστά σαφές ότι η επιλογή μοντέλου εξαρτάται από τις απαιτήσεις της εκάστοτε εφαρμογής.

5.5 Συζήτηση και ερμηνεία αποτελεσμάτων

Καταρχάς, η ερμηνεία των πειραματικών αποτελεσμάτων υποδεικνύει ότι η υπεροχή του MMS_TTS_VITS δεν αποτελεί τυχαίο εύρημα, αλλά προκύπτει από θεμελιώδεις

αρχιτεκτονικές επιλογές που επηρεάζουν άμεσα τη διαδικασία σύνθεσης φωνής. Η ενοποίηση των σταδίων acoustic modeling και waveform generation επιτρέπει τη συνεχή ροή πληροφορίας χωρίς απώλειες ή ασυνέχειες, γεγονός που οδηγεί σε πιο συνεκτική και φυσική παραγωγή ομιλίας. Στη συνέχεια, η χαμηλότερη απόδοση του SpeechT5 μπορεί να αποδοθεί σε πολλαπλούς παράγοντες, συμπεριλαμβανομένης της εξάρτησης από εξωτερικά speaker embeddings και της πιθανής ασυμβατότητας μεταξύ των επιμέρους components. Η εμφάνιση σφαλμάτων κατά τη φόρτωση των embeddings και η χρήση fallback μηχανισμών, όπως παρατηρήθηκε κατά την εκτέλεση, ενδέχεται να επηρεάζουν αρνητικά την ποιότητα της παραγόμενης ομιλίας, οδηγώντας σε υποβάθμιση της φυσικότητας. Η ανάλυση των φασματικών χαρακτηριστικών υποδεικνύει ότι η ποιότητα της ομιλίας δεν εξαρτάται αποκλειστικά από τη γενική αρχιτεκτονική του μοντέλου, αλλά και από τη λεπτομερή διαχείριση των συχνοτήτων. Η υπερβολική ενίσχυση υψηλών συχνοτήτων στο SpeechT5 οδηγεί σε ακουστικά artifacts, τα οποία μειώνουν την αντιληπτή ποιότητα, ακόμη και αν άλλες μετρικές εμφανίζονται ικανοποιητικές.

Η διαφορά στην ενεργειακή κατανομή υποδεικνύει ότι η φυσικότητα της ομιλίας σχετίζεται στενά με τη δυναμική του σήματος. Το MMS_TTS_VITS, μέσω της παραγωγής σημάτων με υψηλότερη RMS ενέργεια, επιτυγχάνει πιο “γεμάτη” και ρεαλιστική απόδοση, ενώ το SpeechT5 εμφανίζει περιορισμένη δυναμική, γεγονός που επηρεάζει αρνητικά την εμπειρία ακρόασης. Το παρατηρούμενο trade-off μεταξύ ποιότητας και υπολογιστικής αποδοτικότητας αναδεικνύει ένα από τα βασικά προβλήματα στη σχεδίαση συστημάτων TTS. Ενώ το MMS_TTS_VITS προσφέρει υψηλότερη ποιότητα, το αυξημένο υπολογιστικό κόστος μπορεί να αποτελεί περιοριστικό παράγοντα σε εφαρμογές πραγματικού χρόνου ή σε περιβάλλοντα με περιορισμένους πόρους. Αντιθέτως, το SpeechT5, παρά τη χαμηλότερη ποιότητα, μπορεί να είναι πιο κατάλληλο σε σενάρια όπου η ταχύτητα αποτελεί κρίσιμο παράγοντα.

Η συνολική ερμηνεία των αποτελεσμάτων υπογραμμίζει τη σημασία της επιλογής κατάλληλου μοντέλου με βάση το συγκεκριμένο πλαίσιο χρήσης. Σε εφαρμογές όπως η παραγωγή πολυμεσικού περιεχομένου ή οι εικονικοί βοηθοί, όπου η φυσικότητα είναι καθοριστική, η χρήση μοντέλων τύπου VITS κρίνεται απαραίτητη. Αντιθέτως, σε περιβάλλοντα όπου απαιτείται γρήγορη απόκριση με περιορισμένους πόρους, πιο

ελαφριές αρχιτεκτονικές μπορεί να αποτελούν βιώσιμη εναλλακτική, ακόμη και με κόστος στην ποιότητα. Η ανάλυση αυτή καταδεικνύει ότι η εξέλιξη των συστημάτων σύνθεσης φωνής δεν εξαρτάται αποκλειστικά από την αύξηση της πολυπλοκότητας των μοντέλων, αλλά από τη βελτιστοποίηση της ισορροπίας μεταξύ ποιότητας, αποδοτικότητας και πρακτικής εφαρμοσιμότητας.

Κεφάλαιο 6: Συμπεράσματα και Μελλοντική Εργασία

6.1 Κύρια ευρήματα της εργασίας

Τα αποτελέσματα της παρούσας μελέτης, προκύπτει ότι η αξιοποίηση προεκπαιδευμένων μοντέλων σύνθεσης φωνής επιτρέπει την επίτευξη υψηλής ποιότητας παραγόμενης ομιλίας χωρίς την ανάγκη εκτεταμένης εκπαίδευσης, γεγονός που μεταβάλλει ουσιαστικά το τεχνοοικονομικό προφίλ ανάπτυξης συστημάτων Text-to-Speech. Η εμπειρική ανάλυση κατέδειξε ότι η χρήση μοντέλων βασισμένων σε ενοποιημένες αρχιτεκτονικές, όπως το MMS_TTS_VITS, οδηγεί σε σαφώς ανώτερη ακουστική απόδοση σε σύγκριση με πιο modular προσεγγίσεις, επιβεβαιώνοντας ότι η συνεκτική μοντελοποίηση της σχέσης μεταξύ κειμένου και κυματομορφής αποτελεί κρίσιμο παράγοντα για τη φυσικότητα της ομιλίας.

Αναλυτικότερα, η συγκριτική αξιολόγηση ανέδειξε ότι η ποιότητα της συνθετικής φωνής δεν εξαρτάται αποκλειστικά από την πολυπλοκότητα του μοντέλου, αλλά από τον τρόπο με τον οποίο ενσωματώνονται και αλληλεπιδρούν τα επιμέρους υποσυστήματα. Το MMS_TTS_VITS, μέσω της end-to-end προσέγγισης, πέτυχε υψηλές τιμές αντιληπτής φυσικότητας, διατηρώντας ταυτόχρονα σταθερότητα στην απόδοσή του ανεξαρτήτως εισόδου. Αντιθέτως, το SpeechT5 εμφάνισε περιορισμούς τόσο ως προς τη συνέπεια όσο και ως προς τη φασματική ισορροπία, γεγονός που επηρέασε αρνητικά την ακουστική ποιότητα. Τα αποτελέσματα κατέδειξαν ότι τα φασματικά χαρακτηριστικά του παραγόμενου σήματος αποτελούν αξιόπιστους δείκτες της αντιληπτής ποιότητας της ομιλίας. Συγκεκριμένα, η ισορροπημένη κατανομή ενέργειας στο φάσμα συχνοτήτων, όπως εκφράζεται μέσω χαμηλότερων τιμών spectral centroid και σταθερότερων MFCC χαρακτηριστικών, συνδέεται άμεσα με πιο φυσική και ευχάριστη ακουστική εμπειρία. Αντιθέτως, αποκλίσεις σε αυτά τα χαρακτηριστικά

οδηγούν σε τεχνητά ηχητικά αποτελέσματα, ακόμη και όταν άλλες μετρικές εμφανίζονται ικανοποιητικές.

Η ανάλυση της ενεργειακής συμπεριφοράς ανέδειξε τη σημασία της δυναμικής του σήματος στην αντίληψη της φυσικότητας. Τα μοντέλα που παράγουν σήματα με επαρκή ένταση και διακύμανση εμφανίζουν σαφώς καλύτερη εκφραστικότητα, στοιχείο ιδιαίτερα σημαντικό για εφαρμογές αλληλεπίδρασης ανθρώπου-μηχανής. Η παρατήρηση αυτή υπογραμμίζει ότι η ποιότητα της ομιλίας δεν αποτελεί μόνο ζήτημα ακρίβειας, αλλά και ζήτημα εκφραστικής πληρότητας.

Τα ευρήματα της εργασίας καταδεικνύουν ότι η αποδοτικότητα των μοντέλων δεν διαφοροποιείται δραστικά όταν αξιολογείται μέσω κανονικοποιημένων μετρικών, όπως ο δείκτης Real-Time Factor. Ωστόσο, η υψηλότερη ποιότητα του MMS_TTS_VITS συνοδεύεται από αυξημένο υπολογιστικό κόστος, γεγονός που αναδεικνύει το θεμελιώδες trade-off μεταξύ ποιότητας και αποδοτικότητας. Η ισορροπία αυτή αποτελεί κρίσιμο παράγοντα στη σχεδίαση πρακτικών συστημάτων, ιδιαίτερα σε περιβάλλοντα με περιορισμένους πόρους. Επιπροσθέτως, η χρήση proxy μετρικών αξιολόγησης αποδείχθηκε αποτελεσματική για τη συγκριτική αποτίμηση των μοντέλων, παρέχοντας επαναλήψιμα και ποσοτικοποιήσιμα αποτελέσματα χωρίς την ανάγκη ανθρώπινης παρέμβασης. Ωστόσο, τα ευρήματα υποδεικνύουν ότι οι αυτοματοποιημένες μετρικές δεν μπορούν να αντικαταστήσουν πλήρως την υποκειμενική αξιολόγηση, καθώς η ανθρώπινη αντίληψη της φυσικότητας επηρεάζεται από παράγοντες που δεν αποτυπώνονται πλήρως σε αριθμητικούς δείκτες.

Η δυνατότητα χρήσης προεκπαιδευμένων μοντέλων μέσω πλατφορμών όπως το Hugging Face επιτρέπει την ταχεία ανάπτυξη εφαρμογών υψηλής ποιότητας, μειώνοντας δραστικά το κόστος και τον χρόνο υλοποίησης. Η εξέλιξη αυτή καθιστά τη σύνθεση φωνής προσβάσιμη σε ευρύτερο φάσμα εφαρμογών, ενισχύοντας τη διάδοση της τεχνολογίας. Τα ευρήματα της παρούσας εργασίας επιβεβαιώνουν ότι η επιλογή αρχιτεκτονικής, η ποιότητα των ενδιάμεσων αναπαραστάσεων και η ισορροπία μεταξύ υπολογιστικής αποδοτικότητας και ακουστικής ποιότητας αποτελούν τους βασικούς άξονες που καθορίζουν την απόδοση των σύγχρονων συστημάτων σύνθεσης φωνής.

6.2 Περιορισμοί και ανοιχτά ζητήματα

Παρά την υψηλή απόδοση των προεκπαιδευμένων μοντέλων που αξιολογήθηκαν, η παρούσα εργασία παρουσιάζει συγκεκριμένους περιορισμούς, οι οποίοι επηρεάζουν τόσο την ερμηνεία των αποτελεσμάτων όσο και τη δυνατότητα γενίκευσής τους σε ευρύτερα σενάρια εφαρμογής. Αρχικά, ένας βασικός περιορισμός αφορά τη χρήση περιορισμένου αριθμού εισόδων κειμένου, γεγονός που μειώνει τη στατιστική ισχύ της ανάλυσης και δεν επιτρέπει την πλήρη αποτύπωση της συμπεριφοράς των μοντέλων σε μεγάλη ποικιλία γλωσσικών δομών. Αν και η επιλογή αυτή έγινε σκόπιμα για την απομόνωση της απόδοσης κατά τη φάση του inference, εντούτοις περιορίζει τη δυνατότητα εξαγωγής γενικευμένων συμπερασμάτων για πραγματικές συνθήκες χρήσης. Επιπροσθέτως, η αξιολόγηση βασίστηκε σε proxy μετρικές, οι οποίες, αν και παρέχουν αντικειμενικότητα και επαναληψιμότητα, δεν αντικατοπτρίζουν πλήρως την ανθρώπινη αντίληψη της φυσικότητας και της ευκρίνειας της ομιλίας. Η απουσία υποκειμενικών δοκιμών με πραγματικούς χρήστες αποτελεί σημαντικό περιορισμό, καθώς παράγοντες όπως η εκφραστικότητα, η συναισθηματική απόχρωση και η προσωδία δεν μπορούν να ποσοτικοποιηθούν πλήρως μέσω αριθμητικών δεικτών. Ως εκ τούτου, η αξιολόγηση της ποιότητας παραμένει εν μέρει προσεγγιστική.

Η εκτέλεση των πειραμάτων σε περιβάλλον CPU εισάγει επιπλέον περιορισμούς ως προς την αποτίμηση της υπολογιστικής αποδοτικότητας. Τα αποτελέσματα που σχετίζονται με τον χρόνο παραγωγής και τον δείκτη Real-Time Factor ενδέχεται να διαφοροποιούνται σημαντικά σε περιβάλλοντα GPU ή edge συσκευών, γεγονός που σημαίνει ότι η αξιολόγηση της αποδοτικότητας δεν μπορεί να θεωρηθεί καθολικά αντιπροσωπευτική. Η εξάρτηση της απόδοσης από το hardware υπογραμμίζει την ανάγκη για πολυπαραμετρική αξιολόγηση σε διαφορετικά υπολογιστικά περιβάλλοντα.

Αν και αυτό εξυπηρετεί τον στόχο της μελέτης, δηλαδή την αξιολόγηση out-of-the-box απόδοσης, δεν επιτρέπει την πλήρη διερεύνηση των δυνατοτήτων προσαρμογής των μοντέλων σε συγκεκριμένα φωνητικά ή γλωσσικά χαρακτηριστικά. Συνεπώς, τα αποτελέσματα αντικατοπτρίζουν περισσότερο την αρχική ποιότητα των μοντέλων και λιγότερο τη δυνητική τους απόδοση μετά από βελτιστοποίηση. Επιπλέον, τα

προβλήματα που παρατηρήθηκαν κατά τη φόρτωση εξωτερικών datasets και speaker embeddings, όπως η αποτυχία χρήσης συγκεκριμένων resources και η ανάγκη για fallback μηχανισμούς, αναδεικνύουν ζητήματα συμβατότητας και εξάρτησης από το οικοσύστημα εργαλείων. Τα ζητήματα αυτά δεν επηρεάζουν μόνο την ποιότητα της παραγόμενης ομιλίας, αλλά και την αξιοπιστία της πειραματικής διαδικασίας, καθώς ενδέχεται να εισάγουν μη ελεγχόμενες μεταβλητές.

6.3 Μελλοντικές κατευθύνσεις (multilingual TTS, zero-shot voice cloning, real-time inference)

Η εξέλιξη των συστημάτων σύνθεσης φωνής αναμένεται να κινηθεί προς την κατεύθυνση της αυξημένης ευελιξίας, της μείωσης του υπολογιστικού κόστους και της ενίσχυσης της εκφραστικότητας, με ιδιαίτερη έμφαση σε τεχνολογίες που επιτρέπουν την προσαρμογή σε πραγματικό χρόνο και σε πολυγλωσσικά περιβάλλοντα. Μία από τις πλέον υποσχόμενες κατευθύνσεις αφορά την ανάπτυξη multilingual TTS συστημάτων, τα οποία μπορούν να διαχειρίζονται πολλαπλές γλώσσες εντός ενός ενιαίου μοντέλου. Η προσέγγιση αυτή επιτρέπει τη μεταφορά γνώσης μεταξύ γλωσσών υψηλών και χαμηλών πόρων, μειώνοντας την ανάγκη για μεγάλα datasets και διευρύνοντας το πεδίο εφαρμογών της τεχνολογίας.

Η έννοια του zero-shot voice cloning αποτελεί ένα από τα πλέον δυναμικά ερευνητικά πεδία, καθώς επιτρέπει τη δημιουργία νέων φωνητικών προφίλ με ελάχιστα ή και καθόλου δεδομένα εκπαίδευσης. Η δυνατότητα αυτή βασίζεται σε προηγμένες τεχνικές αναπαράστασης φωνής μέσω embeddings, επιτρέποντας την ταχεία προσαρμογή σε νέα φωνητικά χαρακτηριστικά χωρίς την ανάγκη εκτεταμένου fine-tuning. Ωστόσο, η βελτίωση της ποιότητας και της σταθερότητας αυτών των μοντέλων παραμένει ανοικτό ερευνητικό ζήτημα.

Η επίτευξη πραγματικού χρόνου (real-time inference) με υψηλή ποιότητα αποτελεί έναν από τους βασικούς στόχους της σύγχρονης έρευνας. Η μείωση του latency χωρίς απώλεια φυσικότητας απαιτεί την ανάπτυξη αποδοτικότερων αρχιτεκτονικών, όπως non-autoregressive και diffusion-based μοντέλα, καθώς και τη βελτιστοποίηση των αλγορίθμων inference. Η αξιοποίηση εξειδικευμένων hardware επιταχυντών και edge computing υποδομών αναμένεται να διαδραματίσει καθοριστικό ρόλο σε αυτή την

κατεύθυνση. Ακόμη, η ενσωμάτωση συναισθηματικής και προσωδιακής πληροφορίας στα μοντέλα σύνθεσης φωνής αποτελεί κρίσιμο βήμα για την ανάπτυξη πιο φυσικών και διαδραστικών συστημάτων. Η δυνατότητα ελέγχου χαρακτηριστικών όπως ο τόνος, η ένταση και ο ρυθμός ομιλίας σε πραγματικό χρόνο θα επιτρέψει την ανάπτυξη εφαρμογών υψηλής αλληλεπίδρασης, όπως εικονικοί βοηθοί και συστήματα εκπαίδευσης.

Η δημιουργία νέων benchmark datasets και η ενσωμάτωση ανθρώπινης αξιολόγησης σε συνδυασμό με αυτοματοποιημένα εργαλεία θα συμβάλει στη βελτίωση της αξιοπιστίας των πειραματικών αποτελεσμάτων. Οι μελλοντικές κατευθύνσεις της έρευνας στον τομέα της σύνθεσης φωνής αναμένεται να εστιάσουν στη γεφύρωση του χάσματος μεταξύ ποιότητας και αποδοτικότητας, καθώς και στην ανάπτυξη συστημάτων που μπορούν να λειτουργούν σε πραγματικές συνθήκες με υψηλό βαθμό προσαρμοστικότητας και αξιοπιστίας

Βιβλιογραφία

- Liu, Z. (2023). *Comparative analysis of transfer learning in deep learning text-to-speech models on a few-shot dataset*. arXiv. <https://doi.org/10.48550/arxiv.2310.04982>
- Gao, X., Chen, Y., Yue, X., Tsao, Y., & Chen, N. F. (2024). *TTSslow: Slow down text-to-speech with efficiency robustness evaluations*. IEEE/ACM Transactions on Audio, Speech, and Language Processing. <https://doi.org/10.1109/taslp.2025.3533357>
- Kishore, P. V. S. (2025). *Human-level text-to-speech synthesis using style diffusion*. <https://doi.org/10.22214/ijraset.2025.72614>
- Liu, T. (2025). *Clip-TTS: Contrastive text-content and mel-spectrogram*. arXiv. <https://doi.org/10.48550/arxiv.2502.18889>
- Mehrish, A., Majumder, N., Bhardwaj, R., & Poria, S. (2023). *A review of deep learning techniques for speech processing*. Information Fusion, 99, 101869. <https://doi.org/10.1016/j.inffus.2023.101869>
- Goswami, N., & Harada, T. (2022). *SATTS: Speaker attractor text-to-speech*. arXiv. <https://doi.org/10.48550/arxiv.2207.06011>

- Maniati, G., et al. (2022). *SOMOS: The Samsung open MOS dataset*. Interspeech. <https://doi.org/10.21437/interspeech.2022-10922>
- Chen, L.-W., Watanabe, S., & Rudnicky, A. (2023). *Vector quantized approach for TTS*. arXiv. <https://doi.org/10.48550/arxiv.2302.04215>
- Ma, S., Oh, J., Kim, M., & Kim, J.-H. (2025). *Survey on deep learning-based speech technologies*. KSII Transactions. <https://doi.org/10.3837/tiis.2025.05.002>
- Aghalarov, M., et al. (2025). *Unified TTS framework using VITS*. <https://doi.org/10.25045/jpit.v16.i2.07>
- Yiwere, M., et al. (2022). *TTS pipeline and evaluation methodology*. IEEE Access. <https://doi.org/10.1109/access.2022.3170836>
- Rai, A., et al. (2024). *Tacotron2 and WaveGlow for low-resource TTS*. <https://doi.org/10.3126/kjse.v8i1.69276>
- Mishev, K., et al. (2020). *Deep learning TTS for Macedonian*. Applied Sciences. <https://doi.org/10.3390/app10196882>
- Tu, T., et al. (2020). *Semi-supervised learning for multi-speaker TTS*. Interspeech. <https://doi.org/10.21437/interspeech.2020-1824>
- Liu, R., et al. (2022). *Knowledge transfer for neural TTS training*. IEEE/ACM TASLP. <https://doi.org/10.1109/taslp.2022.3171974>
- oshi, R., & Garera, N. (2023). Code-mixed text-to-speech synthesis under low-resource constraints. *Lecture Notes in Computer Science*, 151–163. https://doi.org/10.1007/978-3-031-48312-7_12
- Bai, R. H., Likhomanenko, T., Zhang, R., Gu, Z., Aldeneh, Z., & Jaitly, N. (2024). dMel: Speech tokenization made simple. *arXiv preprint arXiv:2407.15835*. <https://doi.org/10.48550/arxiv.2407.15835>
- Gaudier, T., Tahon, M., Larcher, A., & Estève, Y. (2024). Automatic voice identification after speech resynthesis using PPG. *Proceedings of Odyssey 2024*, 1–6. <https://doi.org/10.21437/odyssey.2024-2>
- Joshi, R., & Garera, N. (2023). Rapid speaker adaptation in low-resource text-to-speech systems using synthetic data. *arXiv preprint arXiv:2312.01107*. <https://doi.org/10.48550/arxiv.2312.01107>

- Zhao, W., & Yang, Z. (2023). An emotion speech synthesis method based on VITS. *Applied Sciences*, 13(4), Article 2225. <https://doi.org/10.3390/app13042225>
- Galdino, J. C., Matos, A. N., Svartman, F., & Aluisio, S. (2025). The evaluation of prosody in speech synthesis: A systematic review. *Journal of the Brazilian Computer Society*, 31, 466–487. <https://doi.org/10.5753/jbcs.2025.5468>
- Feng, X., & Yoshimoto, A. (2024). Llama-VITS: Enhancing TTS synthesis with semantic awareness. *arXiv preprint arXiv:2404.06714*. <https://doi.org/10.48550/arxiv.2404.06714>
- Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.
- Hu, G., Tan, H., & Li, R. (2024). A mel spectrogram enhancement paradigm based on CWT in speech synthesis. *Proceedings of IALP 2024*, 401–405. <https://doi.org/10.1109/ialp63756.2024.10661192>
- Barakat, H., Turk, O., & Demiroğlu, C. (2024). Deep learning-based expressive speech synthesis: A systematic review. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024, Article 329. <https://doi.org/10.1186/s13636-024-00329-7>
- Latif, S., Zaidi, A., Cuayáhuítl, H., Shamshad, F., Shoukat, M., & Qadir, J. (2023). Transformers in speech processing: A survey. *arXiv preprint arXiv:2303.11607*. <https://doi.org/10.48550/arxiv.2303.11607>
- Katumba, A., Kagumire, S., Nakatumba-Nabende, J., Quinn, J., & Murindanyi, S. (2025). Building text-to-speech models for low-resourced languages from crowdsourced data. *Applied AI Letters*. <https://doi.org/10.1002/ail2.117>
- Schlittenlacher, J., & Baer, T. (2020). Individually amplified text-to-speech. *arXiv preprint arXiv:2012.02174*.
- Kharitonov, E., Vincent, D., Borsos, Z., Marinier, R., Girgin, S., Pietquin, O., Sharifi, M., Tagliasacchi, M., & Zeghidour, N. (2023). Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11, 1703–1718. https://doi.org/10.1162/tacl_a_00618

- Salvi, D., Hosler, B. C., Bestagini, P., Stamm, M., & Tubaro, S. (2023). TIMIT-TTS: A dataset for multimodal synthetic media detection. *IEEE Access*, 11, 50851–50866. <https://doi.org/10.1109/access.2023.3276480>
- Han, C., & Mesgarani, N. (2022). StyleTTS: A style-based generative model for natural and diverse text-to-speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 19, 283–296. <https://doi.org/10.1109/jstsp.2025.3530171>
- Kumar, G. K., Praveen, V., Kumar, P., Khapra, M. M., & Nandakumar, K. (2023). Towards building text-to-speech systems for the next billion users. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. <https://doi.org/10.1109/ICASSP49357.2023.10096069>
- Nguyen, Q. B., Hoai, L. B., Hoc, N., Quyen, D. B., & Phuong, N. T. (2022). Development of Smartcall Vietnamese text-to-speech system. *VNU Journal of Science: Computer Science and Communication Engineering*. <https://doi.org/10.25073/2588-1086/vnucsce.348>
- Wang, S., Alexanderson, S., Gustafson, J., Beskow, J., Henter, G., & Székely, É. (2021). Integrated speech and gesture synthesis. *Proceedings of the 2021 International Conference on Multimodal Interaction*. <https://doi.org/10.1145/3462244.3479914>
- Georgiou, E., Kritsis, K., Paraskevopoulos, G., Katsamanis, A., Katsouros, V., & Potamianos, A. (2023). Regotron: Regularizing the Tacotron2 architecture via monotonic alignment loss. *IEEE Spoken Language Technology Workshop (SLT)*, 977–983. <https://doi.org/10.1109/SLT54892.2023.10023268>
- Zhang, Z., Yang, Q., Wang, D., Huang, P., Cao, Y., Ye, K., & Hao, J. (2023). Mitigating unauthorized speech synthesis for voice protection. *Proceedings of the ACM Workshop on Large AI Systems and Models*, 1–10. <https://doi.org/10.1145/3689217.3690615>
- Siuzdak, H., Dura, P., van Rijn, P., & Jacoby, N. (2022). WavThruVec: Latent speech representation as intermediate features for neural speech synthesis. *Proceedings of Interspeech 2022*, 833–837. <https://doi.org/10.21437/Interspeech.2022-10797>

- Lenglet, M., Perrotin, O., & Bailly, G. (2023). The GIPSA-Lab text-to-speech system for the Blizzard Challenge 2023. *Proceedings of the Blizzard Challenge Workshop*. <https://doi.org/10.21437/blizzard.2023-3>
- Das, U., Islam, M. S., Murad, H., Khan, M., Menon, M. H., & Muntasir, T. (2025). BnVITS: A voice cloning approach for single-speaker text-to-speech. *Discover Applied Sciences*, 7. <https://doi.org/10.1007/s42452-025-07351-0>
- Wang, F., Hsu, P.-C., Liu, D.-R., & Lee, H.-Y. (2022). Universal adaptor: Converting mel-spectrograms between different configurations for speech synthesis. *arXiv preprint arXiv:2204.00170*. <https://doi.org/10.48550/arxiv.2204.00170>
- Irengbam, G. S., Wahengbam, N. S., Khumanthem, L. M., & Oinam, P. (2025). Text-to-speech system for Meitei Mayek script. *arXiv preprint arXiv:2508.06870*. <https://doi.org/10.48550/arxiv.2508.06870>
- Mbonimpa, P. S., Tuyizere, D., Biyabani, A., & Tonguz, O. (2025). Edge-based speech transcription and synthesis for Kinyarwanda and Swahili. *arXiv preprint arXiv:2510.16497*. <https://doi.org/10.48550/arxiv.2510.16497>
- Muhamad, S., & Veisi, H. (2022). End-to-end Kurdish speech synthesis based on transfer learning. *Passer Journal of Basic and Applied Sciences*. <https://doi.org/10.24271/psr.2022.351832.1149>
- Yiwere, M., Bigioi, D., Corcoran, P., & Cucu, H. (2022). A text-to-speech pipeline, evaluation methodology, and fine-tuning results for child speech synthesis. *IEEE Access*, 10, 1–12. <https://doi.org/10.1109/ACCESS.2022.3170836>
- Katumba, A., Kagumire, S., Nakatumba-Nabende, J., Quinn, J., & Murindanyi, S. (2025). Building text-to-speech models for low-resourced languages from crowdsourced data. *Applied AI Letters*. <https://doi.org/10.1002/ail2.117>
- Schlittenlacher, J., & Baer, T. (2020). Individually amplified text-to-speech. *arXiv preprint arXiv:2012.02174*.

Παράρτημα

```
!pip -q install "transformers==4.41.2"
!pip -q install "datasets==2.19.2"
!pip -q install "accelerate>=0.30.0"
!pip -q install soundfile librosa scipy pandas numpy matplotlib
seaborn pesq jiwer tabulate
```

```
import os
import gc
import time
import json
import math
import warnings
from pathlib import Path
```

```
import numpy as np
import pandas as pd
import torch
import soundfile as sf
import librosa
import librosa.display
import matplotlib.pyplot as plt
import seaborn as sns
```

```
from IPython.display import Audio, display
```

```
warnings.filterwarnings("ignore")
```

```
DEVICE = "cuda" if torch.cuda.is_available() else "cpu"
```

```
OUTPUT_DIR = Path("tts_results")
AUDIO_DIR = OUTPUT_DIR / "audio"
PLOTS_DIR = OUTPUT_DIR / "plots"
ADVANCED_PLOTS_DIR = PLOTS_DIR / "advanced_visuals"
REPORT_DIR = OUTPUT_DIR / "report"
```

```
for folder in [OUTPUT_DIR, AUDIO_DIR, PLOTS_DIR,
               ADVANCED_PLOTS_DIR, REPORT_DIR]:
    folder.mkdir(parents=True, exist_ok=True)
```

```
sns.set_theme(style="whitegrid", context="talk")
```

```
print("Device:", DEVICE)
print("Output directory:", OUTPUT_DIR)
```

```

TEXTS = [
    "Artificial intelligence is transforming the way humans
    interact with technology.",
    "Text to speech systems convert written language into natural
    sounding speech.",
    "Pretrained models from Hugging Face can be used for speech
    synthesis experiments.",
    "Modern neural speech synthesis models can produce expressive
    and realistic audio.",
    "This experiment compares pretrained text to speech systems
    using objective metrics."
]

```

```

REFERENCE_AUDIO = None
# Αν υπάρχει πραγματικό reference audio, ανεβάζει το στο Colab
και βάζουμε:
# REFERENCE_AUDIO = "reference.wav"

```

```

print("Number of texts:", len(TEXTS))

```

```

def clean_memory():
    gc.collect()
    if torch.cuda.is_available():
        torch.cuda.empty_cache()

```

```

def save_audio(path, waveform, sample_rate):
    waveform = np.asarray(waveform).squeeze()

    if waveform.ndim > 1:
        waveform = waveform.mean(axis=0)

    waveform = waveform.astype(np.float32)

    max_abs = np.max(np.abs(waveform)) if len(waveform) else 0
    if max_abs > 1.0:
        waveform = waveform / max_abs

    sf.write(str(path), waveform, int(sample_rate))

```

```

def load_audio(path, target_sr=None):
    y, sr = librosa.load(str(path), sr=target_sr, mono=True)
    return y, sr

```

```

def audio_duration(path):
    y, sr = load_audio(path)
    if sr == 0:

```

```

        return np.nan
    return len(y) / sr

def compute_rtf(generation_time, audio_path):
    duration = audio_duration(audio_path)
    if duration == 0 or np.isnan(duration):
        return np.nan
    return generation_time / duration

def compute_basic_audio_features(audio_path):
    y, sr = load_audio(audio_path)

    duration = len(y) / sr
    rms = float(np.mean(librosa.feature.rms(y=y)))
    zcr = float(np.mean(librosa.feature.zero_crossing_rate(y)))
    centroid =
float(np.mean(librosa.feature.spectral_centroid(y=y, sr=sr)))
    bandwidth =
float(np.mean(librosa.feature.spectral_bandwidth(y=y, sr=sr)))
    rolloff = float(np.mean(librosa.feature.spectral_rolloff(y=y,
sr=sr)))
    flatness =
float(np.mean(librosa.feature.spectral_flatness(y=y)))

    mfcc = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=13)
    mfcc_mean = float(np.mean(mfcc))
    mfcc_std = float(np.std(mfcc))

    return {
        "duration_sec": duration,
        "sample_rate": sr,
        "rms_energy": rms,
        "zero_crossing_rate": zcr,
        "spectral_centroid": centroid,
        "spectral_bandwidth": bandwidth,
        "spectral_rolloff": rolloff,
        "spectral_flatness": flatness,
        "mfcc_mean": mfcc_mean,
        "mfcc_std": mfcc_std,
    }

def compute_mcd_between_files(reference_path, generated_path,
n_mfcc=13):
    try:
        ref, _ = load_audio(reference_path, target_sr=22050)
        gen, _ = load_audio(generated_path, target_sr=22050)

```

```

min_len = min(len(ref), len(gen))
ref = ref[:min_len]
gen = gen[:min_len]

if min_len < 22050:
    return np.nan

ref_mfcc = librosa.feature.mfcc(y=ref, sr=22050,
n_mfcc=n_mfcc)
gen_mfcc = librosa.feature.mfcc(y=gen, sr=22050,
n_mfcc=n_mfcc)

min_frames = min(ref_mfcc.shape[1], gen_mfcc.shape[1])
ref_mfcc = ref_mfcc[:, :min_frames]
gen_mfcc = gen_mfcc[:, :min_frames]

diff = ref_mfcc - gen_mfcc
mcd = (10.0 / np.log(10)) * np.sqrt(2 *
np.mean(np.sum(diff ** 2, axis=0)))
    return float(mcd)
except Exception:
    return np.nan

def compute_pesq_between_files(reference_path, generated_path):
    try:
        from pesq import pesq

        ref, _ = load_audio(reference_path, target_sr=16000)
        gen, _ = load_audio(generated_path, target_sr=16000)

        min_len = min(len(ref), len(gen))
        ref = ref[:min_len]
        gen = gen[:min_len]

        if min_len < 16000:
            return np.nan

        return float(pesq(16000, ref, gen, "wb"))
    except Exception:
        return np.nan

def proxy_mos_score(features, rtf):
    rms = features["rms_energy"]
    zcr = features["zero_crossing_rate"]
    centroid = features["spectral_centroid"]

```

```

    energy_score = min(rms / 0.08, 1.0)
    zcr_score = 1.0 - min(abs(zcr - 0.08) / 0.20, 1.0)
    centroid_score = 1.0 - min(abs(centroid - 1800.0) / 3500.0,
1.0)
    speed_score = 1.0 - min(rtf / 4.0, 1.0)

    score = 1.0 + 4.0 * (
        0.35 * energy_score +
        0.25 * zcr_score +
        0.20 * centroid_score +
        0.20 * speed_score
    )

    return float(np.clip(score, 1.0, 5.0))

def plot_waveform(audio_path, title, save_path):
    y, sr = load_audio(audio_path)

    plt.figure(figsize=(13, 4))
    librosa.display.waveshow(y, sr=sr)
    plt.title(title, fontsize=16, weight="bold")
    plt.xlabel("Time (s)")
    plt.ylabel("Amplitude")
    plt.tight_layout()
    plt.savefig(save_path, dpi=300, bbox_inches="tight")
    plt.close()

def plot_spectrogram(audio_path, title, save_path):
    y, sr = load_audio(audio_path)

    mel = librosa.feature.melspectrogram(y=y, sr=sr, n_mels=80)
    mel_db = librosa.power_to_db(mel, ref=np.max)

    plt.figure(figsize=(13, 4))
    librosa.display.specshow(mel_db, sr=sr, x_axis="time",
y_axis="mel", cmap="magma")
    plt.colorbar(format="+2.0f dB")
    plt.title(title, fontsize=16, weight="bold")
    plt.tight_layout()
    plt.savefig(save_path, dpi=300, bbox_inches="tight")
    plt.close()

def synthesize_with_mms_tts(text, output_path):
    from transformers import VitsModel, AutoTokenizer

```

```

model_name = "facebook/mms-tts-eng"

tokenizer = AutoTokenizer.from_pretrained(model_name)
model = VitsModel.from_pretrained(model_name).to(DEVICE)
model.eval()

inputs = tokenizer(text, return_tensors="pt").to(DEVICE)

start = time.time()
with torch.no_grad():
    output = model(**inputs).waveform
elapsed = time.time() - start

waveform = output.detach().cpu().numpy().squeeze()
sample_rate = model.config.sampling_rate

save_audio(output_path, waveform, sample_rate)

del model, tokenizer
clean_memory()

return elapsed, sample_rate

def get_speecht5_speaker_embedding():
    """
    Stable solution:
    Try loading CMU Arctic xvectors with compatible datasets
    version.
    If it fails, use a deterministic normalized random xvector.
    SpeechT5 requires a 512-dimensional speaker embedding.
    """
    try:
        from datasets import load_dataset

        ds = load_dataset(
            "Matthijs/cmu-arctic-xvectors",
            split="validation",
            trust_remote_code=True
        )

        emb = torch.tensor(ds[7306]["xvector"],
dtype=torch.float32).unsqueeze(0)
        return emb
    except Exception as e:
        print("Speaker embedding dataset failed. Using fallback
embedding.")
        print("Reason:", str(e))

```

```

        rng = np.random.default_rng(42)
        emb = rng.normal(0, 1, size=(1, 512)).astype(np.float32)
        emb = emb / np.linalg.norm(emb, axis=1, keepdims=True)
        return torch.tensor(emb, dtype=torch.float32)

def synthesize_with_speecht5(text, output_path):
    from transformers import SpeechT5Processor,
    SpeechT5ForTextToSpeech, SpeechT5HifiGan

    processor =
    SpeechT5Processor.from_pretrained("microsoft/speecht5_tts")
    model =
    SpeechT5ForTextToSpeech.from_pretrained("microsoft/speecht5_tts")
    .to(DEVICE)
    vocoder =
    SpeechT5HifiGan.from_pretrained("microsoft/speecht5_hifigan").to(
    DEVICE)

    model.eval()
    vocoder.eval()

    speaker_embeddings =
    get_speecht5_speaker_embedding().to(DEVICE)
    inputs = processor(text=text, return_tensors="pt").to(DEVICE)

    start = time.time()
    with torch.no_grad():
        speech = model.generate_speech(
            inputs["input_ids"],
            speaker_embeddings,
            vocoder=vocoder
        )
    elapsed = time.time() - start

    waveform = speech.detach().cpu().numpy()
    sample_rate = 16000

    save_audio(output_path, waveform, sample_rate)

    del processor, model, vocoder
    clean_memory()

    return elapsed, sample_rate

MODELS = {
    "MMS_TTS_VITS": synthesize_with_mms_tts,
    "SpeechT5": synthesize_with_speecht5,
}

```

```

print("Models:")
for model_name in MODELS:
    print("-", model_name)

results = []

for model_name, synth_func in MODELS.items():
    print("\n" + "=" * 70)
    print("Running model:", model_name)
    print("=" * 70)

    for idx, text in enumerate(TEXTS, start=1):
        safe_model = model_name.replace("/", "_")
        audio_path = AUDIO_DIR / f"{safe_model}_text_{idx}.wav"

        try:
            gen_time, sr = synth_func(text, audio_path)
            rtf = compute_rtf(gen_time, audio_path)
            features = compute_basic_audio_features(audio_path)
            mos_proxy = proxy_mos_score(features, rtf)

            if REFERENCE_AUDIO is not None:
                pesq_score =
compute_pesq_between_files(REFERENCE_AUDIO, audio_path)
                mcd_score =
compute_mcd_between_files(REFERENCE_AUDIO, audio_path)
            else:
                pesq_score = np.nan
                mcd_score = np.nan

            row = {
                "model": model_name,
                "text_id": idx,
                "text": text,
                "audio_path": str(audio_path),
                "generation_time_sec": gen_time,
                "rtf": rtf,
                "proxy_mos": mos_proxy,
                "pesq": pesq_score,
                "mcd": mcd_score,
                **features,
            }

            results.append(row)

        print(f"✓ Text {idx}")

```

```

        print(f"  Audio: {audio_path}")
        print(f"  Generation time: {gen_time:.3f} sec")
        print(f"  RTF: {rtf:.3f}")
        print(f"  Proxy MOS: {mos_proxy:.3f}")

    except Exception as e:
        print(f"  Failed: {model_name} | Text {idx}")
        print("Error:", str(e))

df = pd.DataFrame(results)
df.to_csv(OUTPUT_DIR / "tts_evaluation_results.csv", index=False)

print("\nExperiment finished.")
display(df)

for _, row in df.iterrows():
    print("\n" + "-" * 70)
    print(f"Model: {row['model']} | Text {row['text_id']}")
    print(row["text"])
    display(Audio(row["audio_path"]))

for _, row in df.iterrows():
    model = row["model"]
    text_id = row["text_id"]
    audio_path = row["audio_path"]

    wave_path = PLOTS_DIR /
f"{model}_text_{text_id}_waveform.png"
    spec_path = PLOTS_DIR /
f"{model}_text_{text_id}_spectrogram.png"

    plot_waveform(
        audio_path,
        title=f"Waveform - {model} - Text {text_id}",
        save_path=wave_path
    )

    plot_spectrogram(
        audio_path,
        title=f"Mel Spectrogram - {model} - Text {text_id}",
        save_path=spec_path
    )

print("Basic plots saved in:", PLOTS_DIR)

summary = df.groupby("model").agg({
    "generation_time_sec": "mean",

```

```

        "rtf": "mean",
        "proxy_mos": "mean",
        "duration_sec": "mean",
        "rms_energy": "mean",
        "zero_crossing_rate": "mean",
        "spectral_centroid": "mean",
        "spectral_bandwidth": "mean",
        "spectral_rolloff": "mean",
        "spectral_flatness": "mean",
        "mfcc_mean": "mean",
        "mfcc_std": "mean",
        "pesq": "mean",
        "mcd": "mean",
    }).reset_index()

summary.to_csv(OUTPUT_DIR / "tts_model_summary.csv", index=False)

display(summary)

def normalize_metric(series, higher_is_better=True):
    series = series.astype(float)

    if series.max() == series.min():
        return pd.Series([1.0] * len(series), index=series.index)

    norm = (series - series.min()) / (series.max() -
series.min())

    if not higher_is_better:
        norm = 1.0 - norm

    return norm

ranking = summary.copy()

ranking["norm_proxy_mos"] =
normalize_metric(ranking["proxy_mos"], True)
ranking["norm_rtf"] = normalize_metric(ranking["rtf"], False)
ranking["norm_generation_time"] =
normalize_metric(ranking["generation_time_sec"], False)
ranking["norm_rms_energy"] =
normalize_metric(ranking["rms_energy"], True)
ranking["norm_spectral_flatness"] =
normalize_metric(ranking["spectral_flatness"], False)

ranking["final_score"] = (
    0.45 * ranking["norm_proxy_mos"] +
    0.20 * ranking["norm_rtf"] +

```

```

    0.15 * ranking["norm_generation_time"] +
    0.10 * ranking["norm_rms_energy"] +
    0.10 * ranking["norm_spectral_flatness"]
)

ranking = ranking.sort_values("final_score", ascending=False)
ranking.to_csv(OUTPUT_DIR / "tts_model_ranking.csv", index=False)

display(ranking)

def thesis_barplot(data, metric, title, ylabel, filename,
higher_is_better=True):
    if metric not in data.columns:
        return

    plot_data = data.sort_values(metric, ascending=not
higher_is_better)

    plt.figure(figsize=(11, 6))
    ax = sns.barplot(data=plot_data, x="model", y=metric,
palette="viridis")

    for container in ax.containers:
        ax.bar_label(container, fmt="%.3f", fontsize=11)

    plt.title(title, fontsize=18, weight="bold")
    plt.xlabel("TTS Model", fontsize=14)
    plt.ylabel(ylabel, fontsize=14)
    plt.xticks(rotation=20, ha="right")
    plt.tight_layout()
    plt.savefig(ADVANCED_PLOTS_DIR / filename, dpi=300,
bbox_inches="tight")
    plt.show()

metrics_to_plot = [
    ("proxy_mos", "Average Proxy MOS per Model", "Proxy MOS",
"01_proxy_mos.png", True),
    ("generation_time_sec", "Average Generation Time per Model",
"Seconds", "02_generation_time.png", False),
    ("rtf", "Average Real Time Factor per Model", "RTF",
"03_rtf.png", False),
    ("duration_sec", "Average Audio Duration per Model",
"Seconds", "04_duration.png", True),
    ("rms_energy", "Average RMS Energy per Model", "RMS Energy",
"05_rms_energy.png", True),
    ("zero_crossing_rate", "Average Zero Crossing Rate per
Model", "ZCR", "06_zcr.png", False),

```

```

        ("spectral_centroid", "Average Spectral Centroid per Model",
"Hz", "07_spectral_centroid.png", False),
        ("spectral_bandwidth", "Average Spectral Bandwidth per
Model", "Hz", "08_spectral_bandwidth.png", False),
        ("spectral_flatness", "Average Spectral Flatness per Model",
"Flatness", "09_spectral_flatness.png", False),
        ("mfcc_std", "Average MFCC Standard Deviation per Model",
"MFCC STD", "10_mfcc_std.png", True),
    ]

for metric, title, ylabel, filename, higher in metrics_to_plot:
    thesis_barplot(summary, metric, title, ylabel, filename,
higher)

radar_df = summary.copy()

radar_metrics = {
    "proxy_mos": True,
    "generation_time_sec": False,
    "rtf": False,
    "rms_energy": True,
    "zero_crossing_rate": False,
    "spectral_centroid": False,
}

available_radar_metrics = [m for m in radar_metrics if m in
radar_df.columns]

for metric in available_radar_metrics:
    radar_df[f"norm_{metric}"] = normalize_metric(
        radar_df[metric],
        higher_is_better=radar_metrics[metric]
    )

labels = [m.replace("_", " ").title() for m in
available_radar_metrics]
angles = np.linspace(0, 2 * np.pi, len(labels),
endpoint=False).tolist()
angles += angles[:1]

plt.figure(figsize=(9, 9))
ax = plt.subplot(111, polar=True)

for _, row in radar_df.iterrows():
    values = [row[f"norm_{m}"] for m in available_radar_metrics]
    values += values[:1]

    ax.plot(angles, values, linewidth=2.5, label=row["model"])

```

```

ax.fill(angles, values, alpha=0.15)

ax.set_xticks(angles[:-1])
ax.set_xticklabels(labels, fontsize=11)
ax.set_yticks([0.2, 0.4, 0.6, 0.8, 1.0])
ax.set_ylim(0, 1)
ax.set_title("Radar Comparison of TTS Models", fontsize=18,
weight="bold", pad=25)
ax.legend(loc="upper right", bbox_to_anchor=(1.35, 1.10))

plt.tight_layout()
plt.savefig(ADVANCED_PLOTS_DIR / "11_radar_model_comparison.png",
dpi=300, bbox_inches="tight")
plt.show()

heatmap_cols = [f"norm_{m}" for m in available_radar_metrics]
heatmap_df = radar_df[["model"] + heatmap_cols].copy()
heatmap_df = heatmap_df.set_index("model")
heatmap_df.columns = [
    col.replace("norm_", "").replace("_", " ").title()
    for col in heatmap_df.columns
]

plt.figure(figsize=(12, 5))
sns.heatmap(
    heatmap_df,
    annot=True,
    cmap="YlGnBu",
    fmt=".2f",
    linewidths=0.5,
    cbar_kws={"label": "Normalized Score"}
)

plt.title("Normalized Metric Heatmap", fontsize=18,
weight="bold")
plt.xlabel("Metric")
plt.ylabel("Model")
plt.tight_layout()
plt.savefig(ADVANCED_PLOTS_DIR /
"12_normalized_metric_heatmap.png", dpi=300, bbox_inches="tight")
plt.show()

plt.figure(figsize=(11, 6))
ranking_plot = ranking.sort_values("final_score", ascending=True)

plt.barh(
    ranking_plot["model"],
    ranking_plot["final_score"],

```

```

        color=sns.color_palette("mako", len(ranking_plot))
    )

    for i, value in enumerate(ranking_plot["final_score"]):
        plt.text(value + 0.01, i, f"{value:.3f}", va="center",
        fontsize=12)

plt.title("Final Weighted Ranking of TTS Models", fontsize=18,
weight="bold")
plt.xlabel("Final Score")
plt.ylabel("Model")
plt.xlim(0, max(ranking_plot["final_score"]) + 0.15)
plt.tight_layout()
plt.savefig(ADVANCED_PLOTS_DIR / "13_final_weighted_ranking.png",
dpi=300, bbox_inches="tight")
plt.show()

def waveform_grid(dataframe, max_files=10):
    sample_df = dataframe.head(max_files)

    rows = len(sample_df)
    fig, axes = plt.subplots(rows, 1, figsize=(14, 3.0 * rows))

    if rows == 1:
        axes = [axes]

    for ax, (_, row) in zip(axes, sample_df.iterrows()):
        y, sr = librosa.load(row["audio_path"], sr=None)

        librosa.display.waveshow(y, sr=sr, ax=ax)
        ax.set_title(
            f"Waveform - {row['model']} - Text {row['text_id']}",
            fontsize=14,
            weight="bold"
        )
        ax.set_xlabel("Time (s)")
        ax.set_ylabel("Amplitude")

    plt.tight_layout()
    plt.savefig(ADVANCED_PLOTS_DIR / "14_waveform_grid.png",
dpi=300, bbox_inches="tight")
    plt.show()

waveform_grid(df)

def spectrogram_grid(dataframe, max_files=10):
    sample_df = dataframe.head(max_files)

```

```

rows = len(sample_df)
fig, axes = plt.subplots(rows, 1, figsize=(14, 3.5 * rows))

if rows == 1:
    axes = [axes]

for ax, (_, row) in zip(axes, sample_df.iterrows()):
    y, sr = librosa.load(row["audio_path"], sr=None)

    mel = librosa.feature.melspectrogram(y=y, sr=sr,
n_mels=80)
    mel_db = librosa.power_to_db(mel, ref=np.max)

    img = librosa.display.specshow(
        mel_db,
        sr=sr,
        x_axis="time",
        y_axis="mel",
        ax=ax,
        cmap="magma"
    )

    ax.set_title(
        f"Mel Spectrogram - {row['model']} - Text
{row['text_id']}",
        fontsize=14,
        weight="bold"
    )

    fig.colorbar(img, ax=ax, format="%+2.0f dB")

plt.tight_layout()
plt.savefig(ADVANCED_PLOTS_DIR /
"15_mel_spectrogram_grid.png", dpi=300, bbox_inches="tight")
plt.show()

```

```
spectrogram_grid(df)
```

```

distribution_metrics = [
    "generation_time_sec",
    "rtf",
    "proxy_mos",
    "duration_sec",
    "rms_energy",
    "zero_crossing_rate",
    "spectral_centroid",
    "spectral_flatness",
]

```

```

for metric in distribution_metrics:
    if metric not in df.columns:
        continue

    plt.figure(figsize=(10, 6))
    sns.boxplot(data=df, x="model", y=metric, palette="Set2")
    sns.stripplot(data=df, x="model", y=metric, color="black",
alpha=0.65)

    plt.title(f"Distribution of {metric.replace('_', ' '
').title()}", fontsize=18, weight="bold")
    plt.xlabel("Model")
    plt.ylabel(metric.replace("_", " ").title())
    plt.xticks(rotation=20, ha="right")
    plt.tight_layout()
    plt.savefig(
        ADVANCED_PLOTS_DIR / f"16_distribution_{metric}.png",
        dpi=300,
        bbox_inches="tight"
    )
    plt.show()

numeric_cols =
df.select_dtypes(include=[np.number]).columns.tolist()

if len(numeric_cols) >= 2:
    corr = df[numeric_cols].corr()

    plt.figure(figsize=(12, 9))
    sns.heatmap(
        corr,
        annot=True,
        cmap="coolwarm",
        fmt=".2f",
        linewidths=0.5,
        center=0
    )

    plt.title("Correlation Heatmap of Audio and Evaluation
Metrics", fontsize=18, weight="bold")
    plt.tight_layout()
    plt.savefig(ADVANCED_PLOTS_DIR /
"17_correlation_heatmap.png", dpi=300, bbox_inches="tight")
    plt.show()

visual_index = []

```

```

for image_path in sorted(ADVANCED_PLOTS_DIR.glob("*.png")):
    visual_index.append({
        "filename": image_path.name,
        "path": str(image_path),
        "description": image_path.stem.replace("_", " ").title()
    })

visual_index_df = pd.DataFrame(visual_index)
visual_index_df.to_csv(ADVANCED_PLOTS_DIR / "visual_index.csv",
index=False)

display(visual_index_df)

best_model = ranking.iloc[0]["model"] if len(ranking) else "N/A"

report = f"""
# Automatic Text-to-Speech Evaluation Report

## Thesis Topic

Speech synthesis using pretrained models from the Hugging Face
platform.

## Experimental Setup

The experiment evaluates pretrained Text-to-Speech systems using
the same input texts.

## Models Tested

{chr(10).join("- " + model_name for model_name in MODELS.keys())}

## Number of Text Samples

{len(TEXTS)}

## Device

{DEVICE}

## Evaluation Metrics

The following objective and proxy metrics were computed:

1. Generation Time
2. Real Time Factor
3. Audio Duration
4. RMS Energy

```

```

5. Zero Crossing Rate
6. Spectral Centroid
7. Spectral Bandwidth
8. Spectral Rolloff
9. Spectral Flatness
10. MFCC Statistics
11. Proxy MOS
12. PESQ, if reference audio is provided
13. MCD, if reference audio is provided

## Best Model

The best model according to the weighted score is:

**{best_model}**

## Ranking Table

{ranking.to_markdown(index=False)}

## Summary Table

{summary.to_markdown(index=False)}

## Important Note

Proxy MOS is not a true human Mean Opinion Score.
It is an automatic approximation based on acoustic features and
inference efficiency.
For a complete thesis evaluation, human listeners should also
evaluate naturalness and intelligibility.

## Output Files

- Audio files are stored in: `tts_results/audio`
- Plots are stored in: `tts_results/plots`
- Advanced plots are stored in:
  `tts_results/plots/advanced_visuals`
- CSV results are stored in: `tts_results`
"""

report_path = REPORT_DIR / "automatic_tts_report.md"

with open(report_path, "w", encoding="utf-8") as f:
    f.write(report)

print(report)
print("Report saved at:", report_path)

```

```
metadata = {
    "device": DEVICE,
    "models": list(MODELS.keys()),
    "number_of_texts": len(TEXTS),
    "texts": TEXTS,
    "reference_audio": REFERENCE_AUDIO,
    "output_directory": str(OUTPUT_DIR),
}

with open(OUTPUT_DIR / "experiment_metadata.json", "w",
encoding="utf-8") as f:
    json.dump(metadata, f, indent=4, ensure_ascii=False)

print("Metadata saved.")

!zip -r tts_results.zip tts_results > /dev/null

print("DONE.")
print("Download: tts_results.zip")
```