National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics

Text-driven Articulate Talking Face Generation

DIPLOMA THESIS

of

Georgios Milis

Supervisor: Petros Maragos
Professor NTUA

Co-supervisor: Anastasios Roussos
Principal Researcher ICS-FORTH

Computer Vision, Speech Communication and Signal Processing Group
Athens, March 2024

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Computer Vision, Speech Communication and Signal Processing Group

Text-driven Articulate Talking Face Generation

DIPLOMA THESIS

of

Georgios Milis

Supervisor: Petros Maragos
Professor NTUA

Co-supervisor: Anastasios Roussos
Principal Researcher ICS-FORTH

Approved by the Examining Committee on 28th March, 2024.

.....	.....	.....
Petros Maragos	Alexandros Potamianos	Athanasios Rontogiannis
Professor NTUA	Associate Professor NTUA	Associate Professor NTUA

Athens, March 2024

.....

Georgios Milis
Graduate of Electrical and Computer Engineering NTUA

Στην οικογένειά μου

Abstract

Recent advances in deep learning for sequential data have given rise to fast and powerful models that produce realistic videos of talking humans, creating a new era of lifelike virtual experiences. These endeavors not only push the boundaries of audiovisual synthesis but also hold immense potential for applications spanning entertainment, communication, and education. The state of the art in talking face generation focuses mainly on audio-driven methods, which are conditioned on either real or synthetic audios. However, having the ability to directly synthesize talking humans from text transcriptions is particularly beneficial for many applications and is expected to receive more and more attention, following the recent breakthroughs in large language models. A text-driven system can provide an animated avatar that utters a conversational agent's response, paving the way towards a more natural mode of human-machine interaction. Regarding text-driven generation, the predominant approach has been to employ a cascaded 2-stage architecture of a text-to-speech module followed by an audio-driven talking face generator. However this ignores the highly complex interplay between audio and visual streams that occurs during speaking.

In this Diploma Thesis, we construct a text-driven audiovisual speech synthesizer that uses transformers for sequence modeling and does not follow the aforementioned cascaded approach. Instead, our method, which we call NEUral Text to ARticulate Talk (NEUTART), uses joint audiovisual modeling, as well as speech-informed 3D facial reconstructions and various perceptual losses for visual supervision. Notably, we incorporate a lipreading loss which adds realism to the speaker's mouth movements. The proposed model incorporates an audiovisual module that can generate 3D talking head videos with human-like articulation and synced audiovisual streams by design. Then, a photorealistic module leverages the power of generative adversarial networks to convert the 3D talking head into an RGB video. Our experiments on audiovisual datasets as well as in-the-wild videos reveal state-of-the-art generation quality both in terms of objective metrics and human evaluation, especially when assessing the realism of lip articulation. We also showcase the effectiveness of visual supervision for speech synthesis, since our experiments reveal that NEUTART produces more intelligible speech than a similar text-to-speech architecture.

Keywords: Talking Face Generation, Audiovisual Speech Synthesis, Text-to-Visual Speech, Photorealistic Talking Faces, Portrait Videos, Avatars, Multimodal Learning

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή Πέτρο Μαραγκό που μου έδωσε την ευκαιρία να εκπονήσω τη Διπλωματική μου εργασία στο Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων, δείχνοντάς μου εμπιστοσύνη κατά τη διάρκεια της ανάπτυξης και συγγραφής της. Η παρακολούθηση των μαθημάτων του κυρίου Μαραγκού με ενέπνευσε να στοχεύσω στην ενασχόληση με τον τομέα της επεξεργασίας σημάτων, με ιδιαίτερη έμφαση στην επεξεργασία λόγου και εικόνας.

Επιπλέον, θα ήθελα να ευχαριστήσω εγκάρδια τον Δρ. Αναστάσιο Ρούσσο και τον Δρ. Παναγιώτη Φιλντίση για την πολύτιμη βοήθειά τους ως συνεπιβλέποντες της Διπλωματικής εργασίας. Κατά τη διάρκεια της στενής συνεργασίας μας, μου μετέδωσαν ένα μικρό κομμάτι από τις βαθιές επιστημονικές και τεχνικές τους γνώσεις, ενώ παράλληλα μου παρείχαν ανεκτίμητη ψυχολογική και ηθική υποστήριξη. Ο συνεπιβλέπων Παναγιώτης, μαζί με τα υπόλοιπα μέλη του Εργαστηρίου, αξίζουν ένα ιδιαίτερο ευχαριστώ που με φιλοξένησαν στο χώρο εργασίας τους για ένα σύντομο, αλλά πολύ ευχάριστο και παραγωγικό διάστημα.

Τούτη η εργασία, όπως και η ολοκλήρωση των σπουδών μου στο Πολυτεχνείο, δεν θα ήταν δυνατές χωρίς τη συνεχή στήριξη της οικογένειάς μου, στην οποία οφείλω τα μέγιστα. Τέλος, αισθάνομαι βαθιά ευγνωμοσύνη για κάθε φιλία που μου κράτησε και κρατά συντροφιά. Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω προς τον Απόστολο και τον Γιώργο, αφενός για την έντονη ακαδημαϊκή φλόγα, αλλά κυρίως για το ακόμα εντονότερο χιούμορ τους.

Γεώργιος Μίλης
Μάρτιος 2024

Contents

List of Figures

List of Tables

Εκτεταμένη Ελληνική Περίληψη

1 Εισαγωγή

Η σύνθεση φωνής αποσκοπεί στη δημιουργία ανθρώπινης ομιλίας με βάση ένα κείμενο εισόδου, και έχει εισέλθει στη σύγχρονη καθημερινότητα κυρίως μέσω φωνητικών βοηθών. Τούτη η Διπλωματική Εργασία επικεντρώνεται στο πολυτροπικό πρόβλημα της παραγωγής ομιλούντων προσώπων, δηλαδή την οπτικοακουστική σύνθεση ομιλούντων προσώπων, η οποία περιλαμβάνει την παραγωγή φωνής και βίντεο ενός ανθρώπινου χαρακτήρα που μιλά με ρεαλιστικές κινήσεις χειλιών. Τα βαθιά παραγωγικά μοντέλα μπορούν να αντιμετωπίσουν αυτό το πρόβλημα, μοντελοποιώντας την κατανομή των διαφόρων βίντεο ομιλίας που αντιστοιχούν σε μια δεδομένη είσοδο κειμένου.

Βαθιά Μάθηση

Το feedforward ή γραμμικό δίκτυο είναι μια βασική νευρωνική αρχιτεκτονική που αποτελείται από απλούς μετασχηματισμούς βασισμένους σε βελτιστοποιημένες παραμέτρους. Μπορεί να απεικονιστεί σε ένα δίκτυο από νευρώνες, όπου κάθε νευρώνας πολλαπλασιάζει τις εισόδους του με κάποια βάρη, προσθέτει μια σταθερή τιμή, και εφαρμόζει μια μη γραμμική συνάρτηση ενεργοποίησης.

$$z_j = h\left(\sum_i w_{ji}x_i + w_{j0}\right) \tag{1}$$

Με την αύξηση των επιπέδων και των νευρώνων ανά επίπεδο, το δίκτυο μπορεί να μοντελοποιήσει όλο και πιο περίπλοκες κατανομές (Bishop and Nasrabadi 2006).

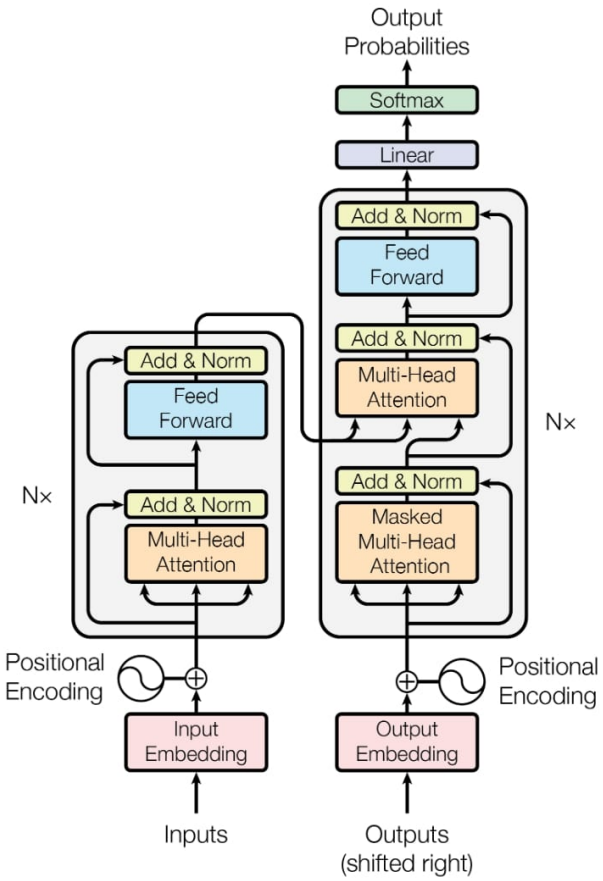
Οι παράμετροι του δικτύου βελτιστοποιούνται επαναληπτικά με την ελαχιστοποίηση συναρτήσεων σφάλματος. Ο αλγόριθμος της κατάβασης κλίσης, ενημερώνει τις παραμέτρους του δικτύου ακολουθώντας το μονοπάτι στον χώρο παραμέτρων που αντιστοιχεί στη μέγιστη τοπική κλίση (LeCun et al. 2015). Έτσι, οι παράμετροι του δικτύου ανανεώνονται με βάση τη σχέση:

$$w^{t+1} = w^t - \eta\frac{\partial\mathcal{L}(w)}{\partial w} \tag{2}$$

Ο αλγόριθμος της αντίστροφης διάδοσης υπολογίζει τις κλίσεις των σφαλμάτων εφαρμόζοντας τον κανόνα της αλυσίδας στον υπολογιστικό γράφο. Λογισμικά βαθιάς μάθησης όπως το PyTorch μπορούν να εκτελέσουν βελτιστοποιήσεις υπολογιστικά, καθώς υλοποιούν αυτόματη διαφόριση συναρτήσεων και διάδοση των κλίσεων.

Παρόλα αυτά, δεδομένα με χρονική διάσταση όπως η ανθρώπινη ομιλία, δεν μπορούν να επεξεργαστούν αποδοτικά από στατικά νευρωνικά δίκτυα, αλλά μοντελοποιούνται από ακολουθιακές αρχιτεκτονικές (Goodfellow et al. 2016). Τα αναδρομικά νευρωνικά δίκτυα (RNNs) έχουν γνωρίσει επιτυχία σε εφαρμογές όπως η επεξεργασία ομιλίας και γλώσσας, αλλά αδυνατούν να διατηρήσουν σχετική πληροφορία για πολύ μακρές ακολουθίες. Τα δίκτυα Long-Term Short-Term Memory (LSTM) επιχειρούν να αντιμετωπίσουν αυτό το πρόβλημα ελέγχοντας τη ροή πληροφορίας στο χρόνο.

Από την άλλη πλευρά, οι μετασχηματιστές (Vaswani et al. 2017) είναι αυτοκωδικοποιητές που επεξεργάζονται ακολουθιακές εισόδους μοντελοποιώντας τις σχέσεις μεταξύ των στοιχείων τους



Σχήμα 1: Η αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή του μετασχηματιστή. Ένας πολυεπίπεδος κωδικοποιητής επεξεργάζεται την ακολουθία εισόδου χρησιμοποιώντας προσοχή πολλών κεφαλών και γραμμικά επίπεδα, δημιουργώντας μια ενδιάμεση αναπαράσταση, η οποία τροφοδοτείται σε κάθε επίπεδο του αποκωδικοποιητή. Η μη επαναληπτική φύση των μετασχηματιστών απαιτεί την έκφραση της διαδοχής των στοιχείων της ακολουθίας, η οποία υλοποιείται προσθέτοντας κωδικοποιήσεις θέσης σε κάθε στοιχείο. Πηγή: Vaswani et al. (2017).

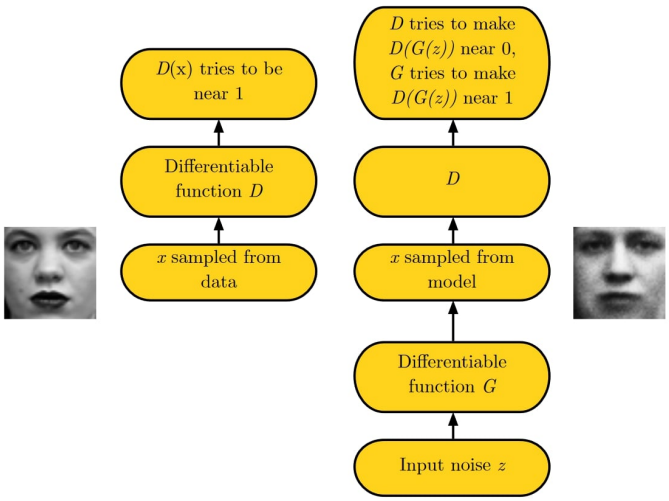
με μηχανισμό προσοχής. Η προσοχή αποσκοπεί στο να αναθεωρεί τη ροή της πληροφορίας σε κάθε επίπεδο, χρησιμοποιώντας σταθμισμένο μέσο όρο με κατάλληλα βάρη που εκφράζουν τη συσχέτιση του τρέχοντος στοιχείου με τα υπόλοιπα στοιχεία της ακολουθίας. Παραθέτουμε την αρχιτεκτονική των μετασχηματιστών στο Σχήμα 1.

Παραγωγικά Μοντέλα

GANs

Τα παραγωγικά δίκτυα με ανταγωνισμό (Generative Adversarial Netowrks ή GANs) αποτελούν σημαντική πρόοδο στη βαθιά παραγωγική μοντελοποίηση, και βασίζονται σε ένα σενάριο της θεωρίας των παιγνίων. Τα GANs περιλαμβάνουν δύο ανταγωνιστικά μοντέλα νευρωνικών δικτύων: ένα γεννητικό δίκτυο $\mathcal{G}$ που μοντελοποιεί την κατανομή των δεδομένων $p(x)$ και ένα διακριτικό δίκτυο $\mathcal{D}$ που εκτιμά την πιθανότητα ότι ένα δείγμα προέρχεται από τα δεδομένα εκπαίδευσης αντί για το $\mathcal{G}$. Η διαδικασία εκπαίδευσης στοχεύει στο να εκπαιδευτεί το $\mathcal{G}$ αρκετά καλά ώστε να μπορεί "εξαπατήσει" το $\mathcal{D}$, παράγοντας ρεαλιστικά δείγματα.

Δεδομένου ότι το $\mathcal{D}$ είναι ισοπίθανο να δει δείγματα από τον γεννήτορα ή το σύνολο δεδομένων, η ισορροπία Nash του παιχνιδιού αντιστοιχεί στο σημείο όπου $\mathcal{G}(z) = p(x)$ και $\mathcal{D}(x) = 0.5$, όπου ο γεννήτορας έχει αποκτήσει τέλεια τη διανομή εκπαίδευσης. Η διαδικασία εκπαίδευσής τους



Σχήμα 2: Οπτικοποίηση της διαδικασίας εκπαίδευσης με ανταγωνισμό σε ένα σύνολο δεδομένων ανθρώπινων προσώπων. Η διαδικασία περιλαμβάνει δύο αντιπάλους που ανταγωνίζονται μεταξύ τους σε ένα παιχνίδι που διαδραματίζεται σε δύο σενάρια. Στο πρώτο σενάριο, παραδείγματα προσώπων x επιλέγονται τυχαία από το σύνολο εκπαίδευσης και χρησιμοποιούνται ως είσοδος για τον εκτιμητή D. Ο στόχος του D είναι να εκτιμήσει την πιθανότητα ότι η είσοδός του είναι πραγματική, το οποίο σημαίνει ότι το $D(x)$ είναι κοντά στο 1. Στο δεύτερο σενάριο, τυχαία z εισέρχονται στον γεννήτορα, και συντίθεται ένα πλαστό δείγμα $G(z)$. Στη συνέχεια, ο εκτιμητής λαμβάνει είσοδο $G(z)$ και εξάγεται το κόστος που υπολογίζεται με $D(G(z)) = 0$ ως επιθυμητή έξοδο, ενώ ο G επιστρέφει αντίστοιχο σφάλμα που υπολογίζεται με $D(G(z)) = 1$ ως επιθυμητή έξοδο. Πηγή: Goodfellow (2016).

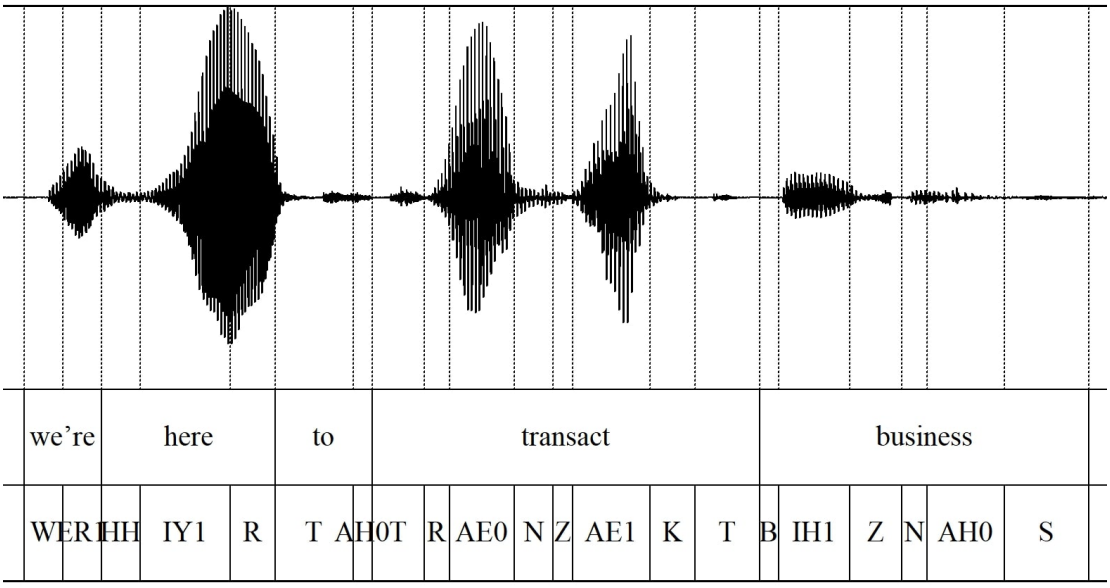
επεξηγείται στο Σχήμα 2.

Τα GANs έχουν χρησιμοποιηθεί ευρέως για τη σύνθεση εικόνων. Το Deep Convolutional GAN (Radford et al. 2015) χρησιμοποιεί αποκλειστικά συνελίξεις για τη μετατροπή ενός τυχαίου διανύσματος θορύβου σε μια εικόνα. Τα StyleGAN και StyleGAN2 (Karras et al. 2020; Karras et al. 2020) χρησιμοποιούν το ανταγωνιστικό πλαίσιο για αυθαίρετη μεταφορά στυλ σε εικόνες, εισάγοντας αρχιτεκτονικές καινοτομίες που προσφέρουν καλύτερο διαχωρισμό των αδρομερών και λεπτομερών χαρακτηριστικών, ενώ επίσης προσφέρουν στοχαστική ποικιλία. Τα GANs έχουν επίσης επιδείξει υπολογιστική αποδοτικότητα στη σύνθεση εικόνων από μια επιγραφή κειμένου, παράγοντας εικόνες υψηλής ανάλυσης σε πολύ λιγότερο χρόνο από τα μοντέλα διάχυσης. Το Generative Interpretable Faces είναι ένα μοντέλο βασισμένο στο StyleGAN2 που δημιουργεί φωτορεαλιστικές εικόνες προσώπων με έλεγχο της γεωμετρίας και του στυλ του προσώπου.

Παρά την επιτυχία τους, τα GANs έχουν μειονεκτήματα, συμπεριλαμβανομένης της πιθανής κατάρρευσης της κατανομής εξόδου τους, η οποία μπορεί να περιορίσει την ποικιλία των δειγμάτων, ή της αστάθειας κατά την εκπαίδευση που μπορεί να οδηγήσει σε μη βέλτιστη σύγκλιση ή απόκλιση.

Μοντέλα Διάχυσης

Οι Sohl-Dickstein et al. (2015) και Ho et al. (2020) πρότειναν έναν νέο τρόπο μάθησης κατανομών, εμπνευσμένο από τη στατιστική φυσική. Διατύπωσαν την ιδέα ότι μια κατανομή δεδομένων μπορεί να μαθευτεί με αργή καταστροφή της δομής της μέσω μιας διαδικασίας θορυβοποίησης των δεδομένων, και εκπαίδευσης ενός μοντέλου με στόχο την αποθορυβοποίησή τους. Τα λεγόμενα *μοντέλα διάχυσης* έχουν χρησιμοποιηθεί ευρύτατα (Yang et al. 2022) για σύνθεση ή συμπλήρωση εικόνων, αποτελώντας τη βάση για μοντέλα σύνθεσης εικόνων με είσοδο κείμενο.



Σχήμα 3: Απεικόνιση της χρονικής ευθυγράμμισης μεταξύ μιας ηχητικής κυματομορφής, του κειμένου απομαγνητοφώνησής της, και της φωνητικής ακολουθίας. Το δείγμα ήχου προέρχεται από το σύνολο δεδομένων TCD-TIMIT (Harte and Gillen 2015).

2 Σύνθεση Φωνής

Φωνητική Μοντελοποίηση

Ένα σύστημα σύνθεσης φωνής (text-to-speech ή TTS) λαμβάνει ως είσοδο ένα απόσπασμα κειμένου ως ακολουθία χαρακτήρων, ενώ η επιθυμητή του έξοδος είναι η κυματομορφή ενός ηχητικού σήματος που μοιάζει με την απόδοση του αρχικού κειμένου από έναν άνθρωπο (Rabiner and Schafer 2010). Προκειμένου να επιτευχθεί αυτό, το κείμενο πρέπει να επεξεργαστεί και να έλθει σε μορφή κατάλληλη προς υπολογισμό.

Το κείμενο χρειάζεται κανονικοποίηση, πράγμα που σημαίνει ότι η κεφαλαία γραφή, τα σύμβολα, και η στίξη είτε αφαιρούνται, είτε αντικαθίστανται κατάλληλα. Στη συνέχεια, το κείμενο πρέπει να αντιστοιχιστεί σε μια ακολουθία *φωνημάτων*, που είναι οι μικρότερες διακριτές μονάδες γλώσσας που μπορούν να χρησιμοποιηθούν για τη σύνθεση λέξεων. Σε αυτήν τη Διπλωματική Εργασία χρησιμοποιήθηκε το ευρέως διαδεδομένο φωνητικό λεξικό του Πανεπιστημίου Carnegie Mellon, το οποίο περιλαμβάνει πάνω από 134.000 αγγλικές λέξεις και τις προφορές τους στο σύνολο ARPAbet που αποτελείται από 39 φωνήματα.

Ευθυγράμμιση

Η εκπαίδευση των μοντέλων TTS γίνεται με τη χρήση ηχητικών κυματομορφών ανθρώπινης φωνής, μαζί με τις απομαγνητοφωνήσεις τους. Κάθε σήμα ήχου πρέπει να ευθυγραμμιστεί χρονικά με τη φωνητική ακολουθία της απομαγνητοφώνησης, πράγμα που σημαίνει ότι το μοντέλο πρέπει να γνωρίζει ακριβώς ποιο τμήμα ήχου περιέχει κάθε φώνημα. Ένα παράδειγμα ευθυγράμμισης μεταξύ ενός σήματος ήχου και της φωνητικής του ακολουθίας δίνεται στο Σχήμα 3.

Το Φασματογράφημα

Το φασματογράφημα περιέχει πληροφορίες σχετικά με το περιεχόμενο συχνοτήτων ενός σήματος καθώς αλλάζει με τον χρόνο. Μπορεί, να υπολογιστεί εφαρμόζοντας τη μετασχηματισμό Fourier σε ένα κινούμενο παράθυρο πάνω στο σήμα, και στη συνέχεια στοιβάζοντας τις εξαγόμενες συχνοτικές αναπαραστάσεις σε μια εικόνα. Τυπικά, το φασματογράφημα $\mathbf{F}$ ενός ψηφιακά δειγματοληπτημένου

σήματος $x[n]$, όπως ένα σήμα ομιλίας, είναι το μέτρο του μετασχηματισμού Fourier βραχέος χρόνου (Short-Time Fourier Transform ή STFT).

$$\mathbf{F}(m,\omega) = |\text{STFT}(m,\omega)|^2 \tag{3}$$

Ο STFT μπορεί να υπολογιστεί χρησιμοποιώντας ένα κινούμενο παράθυρο $w[n]$ που απομονώνει το περιεχόμενο του σήματος σε ένα συγκεκριμένο χρονικό πλαίσιο.

$$\text{STFT}\{x[n]\}(m,\omega) = \sum_n x[n]w[n-m]e^{-j\omega n} \tag{4}$$

Αντί να εκφράσουμε το φασματογράφημα σε φυσική κλίμακα συχνότητας, μπορούμε να μετασχηματίσουμε την συνχνοτική αναπαράσταση από τα Hertz στην κλίμακα mel, η οποία είναι καταλληλότερη για την ανθρώπινη αντίληψη. Η κλίμακα mel είναι μια αντιληπτική κλίμακα των τόνων που κρίνονται από ακροατές πως ισαπέχουν μεταξύ τους, και έχει εξαχθεί μέσα από ψυχοακουστικά πειράματα (Stevens et al. 1937). Προκύπτει πως οι συχνότητες της κλίμακας mel αυξάνουν λογαριθμικά σε σχέση με τις φυσικές συχνότητες σε Hertz.

3 Μοντελοποίηση Προσώπων

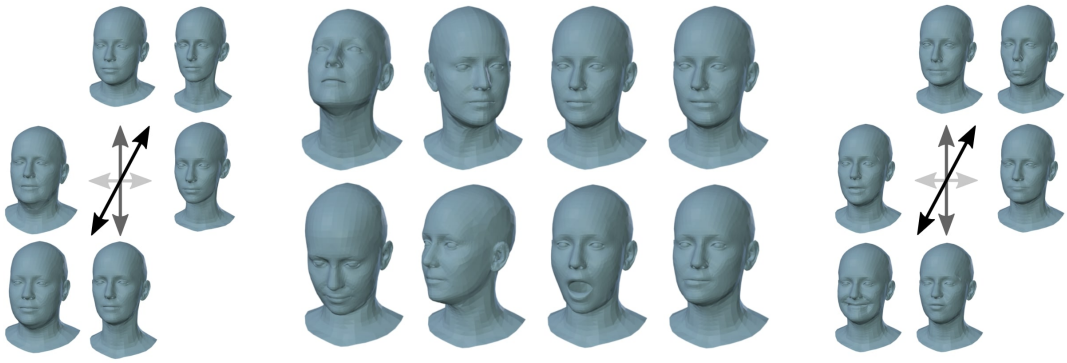
Ένας εξαιρετικά ενδιαφέρων τομέας της όρασης υπολογιστών είναι η μοντελοποίηση προσώπων. Τα μοντέλα προσώπου είναι γνωστά για τις εφαρμογές τους στην ψυχαγωγία, κυρίως στα ειδικά εφέ και τα γραφικά υπολογιστών για παιχνίδια ή ταινίες. Ωστόσο, οι χρήσεις τους εκτείνονται και σε άλλους τομείς. Για παράδειγμα, η αναγνώριση προσώπων μέσω των τρισδιάστατων μοντέλων χρησιμοποιείται σε συστήματα βιομετρικής ασφάλειας ή για εγκληματολογική ανάλυση. Ειδικές εφαρμογές μοντελοποίησης κεφαλών περιλαμβάνουν ιατρική απεικόνιση για χειρουργική κρανίου, γνωσιακή επιστήμη, και νευροεπιστήμες. Πιο πρόσφατες εφαρμογές εντοπίζονται στους τομείς της αλληλεπίδρασης ανθρώπου-μηχανής, της ρομποτικής, και της εικονικής πραγματικότητας.

Πλέγματα

Τα πλέγματα αποτελούνται από N κορυφές στον τρισδιάστατο χώρο, που συνδέονται για να δημιουργήσουν κυρτές πολυγωνικές πλευρές. Τα τριγωνικά πλέγματα είναι ο πιο κοινός τρόπος να αναπαρασταθεί η επιφάνεια που δειγματοληπτείται από τις κορυφές. Η γεωμετρία ενός τρισδιάστατου πλέγματος καθορίζεται από τον πίνακα των κορυφών, ενώ η τοπολογία κωδικοποιείται στη λίστα τριάδων κορυφών που σχηματίζουν κάθε πλευρά. Η διαδικασία μετατροπής ενός αφηρημένου 3Δ γραφικού αντικειμένου σε εικόνα ονομάζεται rasterization και περιλαμβάνει δύο βήματα. Πρώτα εφαρμόζεται ένας μετασχηματισμός ως προς την οπτική γωνία της κάμερας, και έπειτα το 3Δ σχήμα απεικονίζεται στο επίπεδο της εικόνας

Τρισδιάστατα Μορφοποιήσιμα Μοντέλα

Τα *3Δ Μορφοποιήσιμα Μοντέλα* (3ΔΜΜ) είναι στατιστικά μοντέλα τρισδιάστατων σχημάτων που διαχωρίζουν τη μορφή από την εμφάνιση των δεδομένων. Το πρώτο 3ΔΜΜ εισήχθη από τους Blanz and Vetter (1999), οι οποίοι χρησιμοποίησαν υψηλής ποιότητας 3Δ σαρώσεις προσώπων για να εξάγουν τις κύριες συνιστώσες διαφοροποίησής τους. Τα καταγεγραμμένα 3Δ πλέγματα είναι υψηλά συσχετισμένα, καθώς τα χαρακτηριστικά των ανθρώπινων προσώπων είναι αρκετά όμοια. Χρησιμοποιούνται στην επιστήμη της όρασης υπολογιστών και των γραφικών, καθώς επιτρέπουν τη δειγματοληψία τους για τη μοντελοποίηση σχημάτων, καθώς και τη δυνατότητα δημιουργίας νέων πιθανών σχημάτων. Με άλλα λόγια, τα 3ΔΜΜ είναι μια ισχυρή κατανομή πιθανότητας πάνω στη μορφή και τη χρωματική πληροφορία των τρισδιάστατων προσώπων, η οποία μπορεί να χρησιμοποιηθεί σε αλγόριθμους για την ανακατασκευή τρισδιάστατων μοντέλων προσώπων από ελλιπείς πηγές δεδομένων. Επιπλέον, παρέχουν ένα μηχανισμό για τον κωδικοποίηση οποιουδήποτε τρισδιάστατου προσώπου σε ένα χαμηλοδιάστατο χώρο χαρακτηριστικών, γεγονός το οποίο θα αξιοποιήσουμε στη συνέχεια της Διπλωματικής Εργασίας.



Σχήμα 4: Διαχωρισμός των συνιστωσών μεταβολής κεφαλών στο μοντέλο FLAME ως προς το σχήμα, την έκφραση, και την πόζα. Απεικονίζονται οι τρεις πρώτες κύριες συνιστώσες του σχήματος και της έκφρασης σε ±3 τυπικές αποκλίσεις, ενώ παράλληλα φαίνεται και η κίνηση των αρθρώσεων που μοντελοποιούν την πόζα του λαιμού και του σαγονιού. Πηγή: Li et al. (2017).

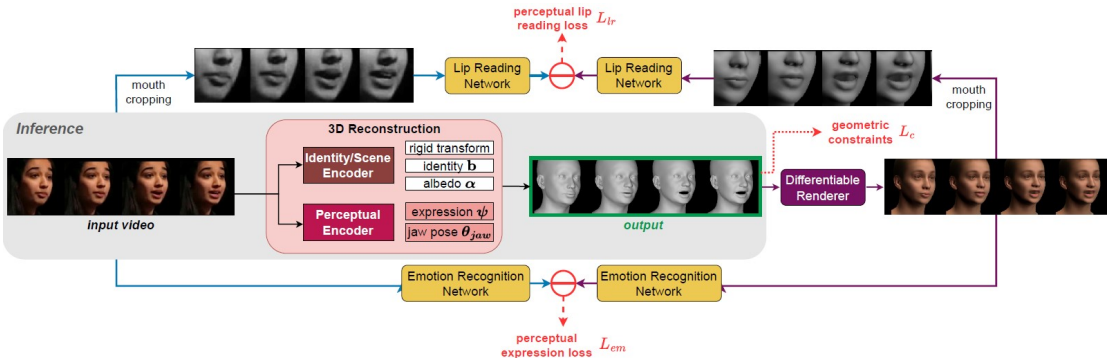
Τα 3ΔΜΜ μαθαίνονται από υψηλής ποιότητας 3Δ σαρώσεις που ευθυγραμμίζονται σε ένα κοινό πλάισιο αναφοράς, ώστε τα σημεία των προσώπων να έλθουν σε αντιστοιχία. Η αντιστοιχία περιλαμβάνει την επαναπαραμετροποίηση των πλεγμάτων σε αναπαράσταση με τον ίδιο αριθμό κορυφών και ίδια τριγωνοποίηση. Έπειτα, τα αντιστοιχισμένα πλέγματα αναλύονται στατιστικά χρησιμοποιώντας την Ανάλυση Κύριων Συνιστωσών (PCA), δημιουργώντας ένα 3Δ μοντέλο ως γραμμική βάση σχημάτων για κάθε συνιστώσα μεταβολής. Εκτός από το σχήμα, η εκφραστικότητα του προσώπου μπορεί να ενσωματωθεί στο μοντέλο προσθέτοντας μια γραμμική βάση εκφράσεων. Απεικονίζουμε την αποσύνθεση των συνιστωσών μεταβολής των προσώπων με ένα παράδειγμα από το μοντέλο FLAME στο Σχήμα 4.

3ΔΜΜ Μέθοδοι

Το Large Scale Facial Model (Booth et al. 2018) είναι ένα 3ΔΜΜ που διαχωρίζει το σχήμα, την έκφραση, τη θέση του κεφαλιού, και την υφή, έχοντας εκπαιδευτεί σε μια μεγάλης κλίμακας βάση δεδομένων. Χρησιμοποιεί τον αλγόριθμο Non-rigid Iterative Closest Point (NICP) για να αποφύγει την ευθυγράμμιση με βάση τα επιφανειακά χαρακτηριστικά του δέρματος και να επιτύχει ευθυγράμμιση ως προς τα ανατομικά χαρακτηριστικά. Το FLAME (Li et al. 2017) αντιστοιχίζει διανύσματα για το σχήμα, την πόζα και την έκφραση σε 3Δ μετατοπίσεις ενός προτύπου πλέγματος. Το μοντέλο επιτυγχάνει καλύτερες ανθρώπινες εκφράσεις μοντελοποιώντας τέσσερις αρθρώσεις με 3Δ περιστροφές: τον λαιμό, το σαγόνι, και τα δύο μάτια. Οι παράμετροι του μοντέλου εκπαιδεύονται για να ελαχιστοποιήσουν το 3Δ σφάλμα ανακατασκευής, ενώ χρησιμοποιούνται βάρη σε θορυβώδεις περιοχές (όπως η κόμη) για βελτιωμένη ακρίβεια. Το MICA (Zielonka et al. 2022) εστιάζει στο ζήτημα της μετρικής ακρίβειας στη μοντελοποίηση ανθρώπινων προσώπων, χρησιμοποιώντας κατάλληλα επισημειωμένα σύνολα δεδομένων. Τα συνελικτικά νευρωνικά δίκτυα (CNNs) αποτελούν μια άλλη προσέγγιση που δεν βασίζεται σε κρυφούς υποχώρους και χρησιμοποιήθηκαν από τους Jackson et al. (2017). Στην κατηγορία των πολυτροπικών μοντέλων, το AVFace (Chatziagapi and Samaras 2023) ενσωματώνει την ακουστική πληροφορία στα 3ΔΜΜ, χρησιμοποιώντας οπτικοακουστικά χαρακτηριστικά για τις παραμέτρους FLAME ανά καρέ από βίντεο προσώπων.

Ανακατασκευή

Το DECA (Feng et al. 2021) είναι ένα μοντέλο τρισδιάστατης ανακατασκευής κεφαλιών με έμφαση στην αποτύπωση λεπτομερειών στην έκφραση. Το EMOCA (Daněček et al. 2022) βασίζεται στο DECA, αλλά δίνει έμφαση στην καλύτερη αποτύπωση των ανθρώπινων εκφράσεων που σχετίζονται με συναισθήματα. Από την άλλη το SPECTRE (Filntisis et al. 2023) αποσκοπεί στην πιστή ανακατασκευή των κινήσεων του στόματος κατά την ομιλία ενός ατόμου. Η μέθοδος χρησι-



Σχήμα 5: Η αρχιτεκτονική και η διαδικασία εκπαίδευσης του SPECTRE. Ένα αρχικό βίντεο τροφοδοτείται στο δίκτυο 3Δ ανακατασκευής, όπου ένας σταθερός κωδικοποιητής βασισμένος στο DECA ανιχνεύει παραμέτρους FLAME και σκηνής (κάμερα, φωτισμός). Στη συνέχεια, ένας κωδικοποιητής στόματος προβλέπει τις ενημερωμένες παραμέτρους έκφρασης και πόζας του σαγονιού, ενώ ένας renderer αποτυπώνει το προβλεπόμενο 3Δ σχήμα σε εικόνα. Η περιοχή του στόματος περικόπτεται τόσο στην είσοδο όσο και στις ανακατασκευασμένες ακολουθίες, και εφαρμόζεται εξαγωγή χαρακτηριστικών από ένα δίκτυο ανάγνωσης χειλιών, προκειμένου να εκτιμηθεί η απόκλιση μεταξύ τους. Το ίδιο γίνεται με ένα αντίστοιχο δίκτυο για την αναγνώριση της έκφρασης του προσώπου, προκειμένου να εκτιμηθεί η απόσταση στον αντιληπτικό χώρο αναπαράστασης εκφράσεων που έχει μάθει το εξωτερικό μοντέλο. Πηγή: Filntisis et al. (2023).

μοποιεί έναν κωδικοποιητή για να εκτιμήσει παραμέτρους του στόματος, συμπεριλαμβανομένης της έκφρασης και της θέσης του σαγονιού, και ένα προεκπαιδευμένο δίκτυο ανάγνωσης χειλιών στο σύνολο δεδομένων LRS3 για να εξάγει διανύσματα χαρακτηριστικών που σχετίζονται με τις κινήσεις κατά την ομιλία, τα οποία προσπαθεί να ταιριάξει μεταξύ αρχικού βίντεο και ανακατασκευής. Παραθέτουμε την αρχιτεκτονική του SPECTRE στο Σχήμα 5, καθώς θα το αξιοποιήσουμε στη συνέχεια.

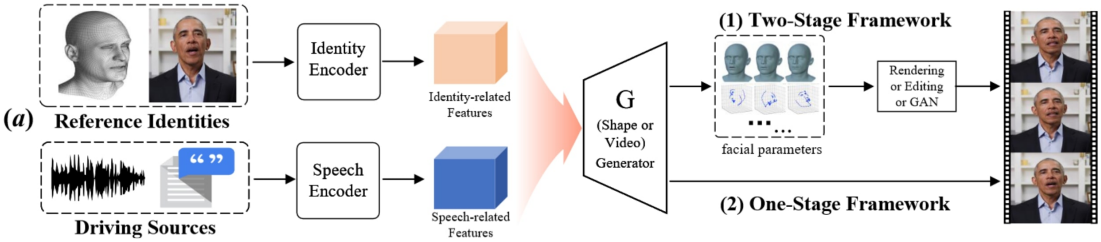
4 Ομιλούντα Πρόσωπα

Η σύνθεση ομιλούντων προσώπων αποσκοπεί στο να παράγει βίντεο με συγχρονισμό και συνέπεια μεταξύ της φωνής και των καρέ του ομιλούντος ατόμου. Η σύνθεση μπορεί να είναι είτε βασισμένη στον ήχο, δηλαδή σύνθεση βίντεο ώστε να ταιριάζει με ένα δεδομένο κομμάτι ομιλίας, είτε βασισμένη στο κείμενο, όπου οι ροές ήχου και βίντεο πρέπει να δημιουργηθούν από ένα τμήμα φυσικής γλώσσας. Σε αυτή τη Διπλωματική Εργασία, αναπτύξαμε ένα μοντέλο βασισμένο στο κείμενο που χρησιμοποιεί μια ενδιάμεση τρισδιάστατη αναπαράσταση. Μια γενική επισκόπηση των συστημάτων δημιουργίας ομιλούντων προσώπων παρουσιάζεται στο Σχήμα 6.

Μοντέλα με Είσοδο Ήχου

Η σύνθεση ομιλούντων προσώπων από ήχο αποσκοπεί στη δημιουργία ενός ρεαλιστικού βίντεο ενός ατόμου που συγχρονίζει την ομιλία του με το εισερχόμενο σήμα φωνής. Τα περισσότερα σύγχρονα μοντέλα ομιλούντων προσώπων δέχονται ήχο ως είσοδο, ενδεχομένως επειδή είναι πιο απλά από τα μοντέλα με είσοδο κειμένου, ενώ καλύπτουν ένα εξίσου ευρύ φάσμα εφαρμογών. Ωστόσο, αναγκάζονται να χρησιμοποιούν μια εξωτερική μέθοδο TTS προκειμένου να είναι σε θέση να δημιουργήσουν πλήρως προσαρμοσμένα βίντεο με αυθαίρετες προτάσεις, κάτι το οποίο περιορίζει την εκφραστικότητα.

Το Neural Voice Puppetry (Thies et al. 2020) είναι ένα deepfake μοντέλο που δημιουργεί φωτορεαλιστικά βίντεο. Χρησιμοποιεί έναν προεκπαιδευμένο δίκτυο για εξαγωγή χαρακτηριστικών ομιλίας, έναν γεννήτορα 3ΔΜΜ ανά καρέ και έναν νευρωνικό renderer. Το VOCA (Cudeiro et al. 2019) εξάγει χαρακτηριστικά ομιλίας από το DeepSpeech και χρησιμοποιεί χρονικές συνελίξεις για να τα μετατρέψει σε μετατοπίσεις 3Δ πλέγματος. Αντίστοιχα, τα μοντέλα των Fan et al. (2022) and



Σχήμα 6: Μια γενική απεικόνιση της διαδικασίας σύνθεσης ομιλούντος προσώπου. Τα εισερχόμενα δεδομένα είναι ένα άτομο αναφοράς, καθώς και μια πηγή για την ομιλία, είτε αυτή είναι ήχος, είτε κείμενο. Για τη σύνθεση φωτορεαλιστικού βίντεο μπορεί να χρησιμοποιηθεί μια ενδιάμεση παραμετρική αναπαράσταση προσώπων. Πηγή: Sheng et al. (2022).

Xing et al. (2023) επιτυγχάνουν πολύ καλά αποτελέσματα στην πρόβλεψη 3Δ πλεγμάτων από ήχο. Το Neural Emotion Director (Paraperas Papantoniou et al. 2022) επεξεργάζεται τις εκφράσεις ενός ατόμου σε βίντεο υπό φυσικές συνθήκες, με είσοδο κάποια συναισθηματική επιγραφή ή κάποιο βίντεο αναφοράς, από το οποίο εξάγει το στυλ των εκφράσεων. Η ενσωμάτωση του συναισθήματος στα ομιλούντα πρόσωπα αποτελεί έναν ενεργό τομέα έρευνας, διευκολυνόμενη από μεγάλα σύνολα δεδομένων όπως το MEAD (Wang et al. 2020).

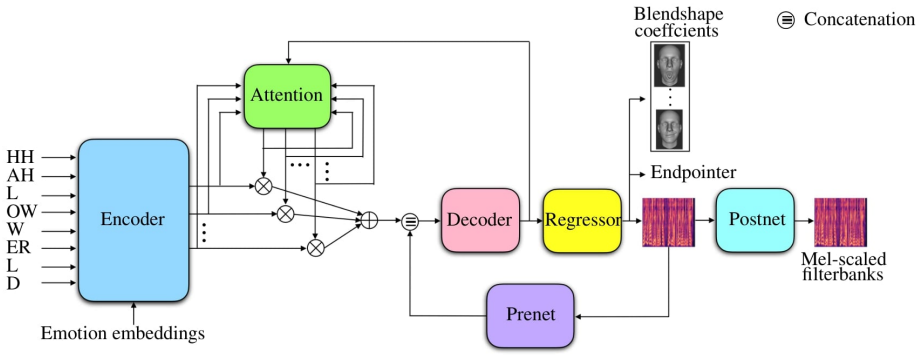
Πολλές μέθοδοι συνθέτουν φωτορεαλιστικά βίντεο χρησιμοποιώντας GANs, όπως το Head2Head (Doukas et al. 2021b). Το Wav2Lip (Prajwal et al. 2020) αντιμετωπίζει το πρόβλημα της σύνθεσης ομιλούντων προσώπων ως δύο υποπροβλήματα: τη σύνθεση εικόνων στόματος, και το συγχρονισμό τους με την είσοδο ήχου. Παρομοίως, τα SadTalker (Zhang et al. 2022c) και VideoReTalking (Cheng et al. 2022) επιτυγχάνουν φωτορεαλιστική few-shot σύνθεση προσώπων με καλό συγχρονισμό στον ήχο.

Μοντέλα με Είσοδο Κειμένου

Η σύνθεση ομιλούντων προσώπων με κείμενο ως είσοδο απαιτεί τη δημιουργία ηχητικής και οπτικής ροής για ένα πρόσωπο με βάση την είσοδο κειμένου, γεγονός που την καθιστά πιο δύσκολη από την προσέγγιση με είσοδο ήχου. Ορισμένες κοινές προσεγγίσεις περιλαμβάνουν τη χρήση ενός μοντέλου TTS εν σειρά με κάποια μέθοδο οδηγούμενη από ήχο, όπως αυτές που περιγράψαμε παραπάνω. Παραδείγματα τέτοιων αρχιτεκτονικών είναι το ObamaNet (Kumar et al. 2017) και το AnyoneNet (Wang et al. 2022). Η ευελιξία της σύνδεσης δύο συστημάτων που διαχωρίζουν το κείμενο, τον ήχο, και το βίντεο έχει οδηγήσει σε αρκετά μοντέλα με παρόμοια αρχιτεκτονική. Ενδεικτικά, τα μοντέλα των Obradović et al. (2022), Song et al. (2022), and Ye et al. (2023) ακολουθούν την εν σειρά αρχιτεκτονική.

Ένα οπτικοακουστικό μοντέλο προτάθηκε από τους Abdelaziz et al. (2021), οι οποίοι επεκτείνουν το TTS μοντέλο Tacotron 2 για σύνθεση ομιλίας (Shen et al. 2018) για να περιλάβει και οπτική πληροφορία, προτείνοντας το μοντέλο AVTacotron2. Η μέθοδός τους παράγει συναισθηματική ομιλία χρησιμοποιώντας διανύσματα συναισθημάτων για να κωδικοποιήσει την απαιτούμενη προσωδία. Τα αντίστοιχα 3Δ πλέγματα και φασματογράμματα δημιουργούνται με αναδρομικό τρόπο. Απεικονίζουμε την αρχιτεκτονική του συστήματος στο Σχήμα 7.

Παρόμοια, το DurIAN (Yu et al. 2019) προσαρμόζει το μοντέλο WaveRNN (Kalchbrenner et al. 2018) για την πρόβλεψη σημείων ενδιαφέροντος γύρω από το πρόσωπο. Το UniFLG (Mitsui et al. 2023) μαθαίνει μια κοινή αναπαράσταση κειμένου και ήχου, επιτρέποντας έτσι τη σύνθεση τόσο με κείμενο όσο και με ήχο. Το οπτικό αποτέλεσμα και των δύο μοντέλων είναι μια ακολουθία σημείων ενδιαφέροντος γύρω από το πρόσωπο και το στόμα. Ωστόσο, αυτή η απλοϊκή μοντελοποίηση δεν είναι κατάλληλη αναπαράσταση για την περίπλοκη άρθρωση των χειλιών, ούτε μπορεί να γενικευθεί σε νέα πρόσωπα.



Σχήμα 7: Το μοντέλο AVTacotron2 για αναδρομική οπτικοακουστική σύνθεση από κείμενο. Μια αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή εξάγει αναπαραστάσεις οπτικής ομιλίας που απεικονίζονται αφενός στο φασματογράφημα, αφετέρου σε παραμέτρους 3Δ προσώπου, ενώ υπάρχει επίσης η δυνατότητα συναισθηματικού ελέγχου. Πηγή: Abdelaziz et al. (2021).

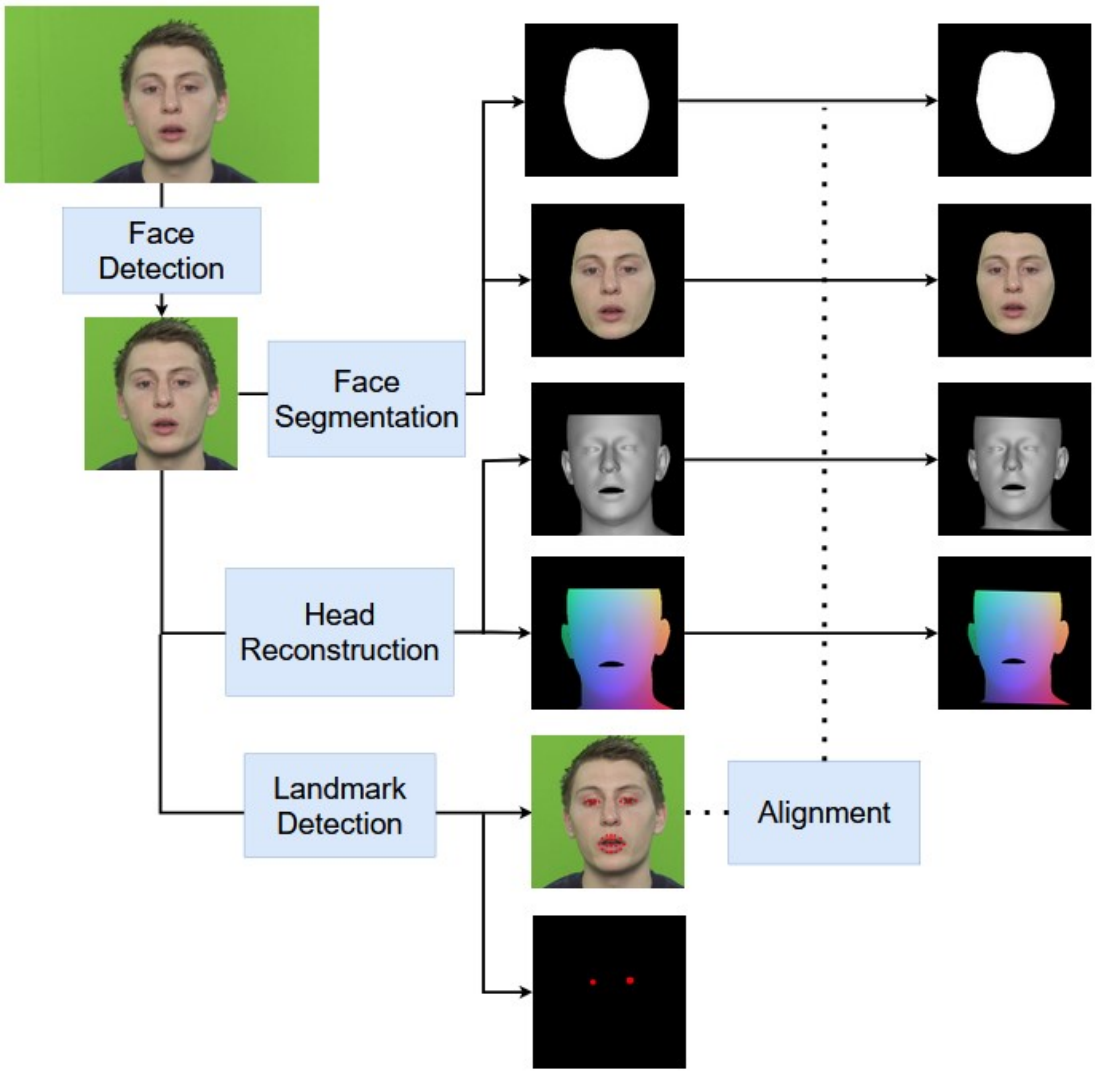
5 Προτεινόμενη Μέθοδος

Σκοπός της παρούσας Διπλωματικής Εργασίας είναι να προτείνει ένα μοντέλο βαθιάς μάθησης για τη δημιουργία βίντεο ομιλούντων προσώπων που εκφέρουν ρεαλιστικά κάποια είσοδο κειμένου. Ενδεικτικές εφαρμογές είναι η δημιουργία χαρακτήρων για εικονικούς βοηθούς, εκπαιδευτικά βίντεο, εργαλεία προσβασιμότητας και διεπαφές ανθρώπου-μηχανής. Έτσι, το μοντέλο επικεντρώνεται στο ρεαλισμό της κίνησης των χειλιών, καθώς και το συγχρονισμό των ακουστικών και οπτικών ροών, λόγω της μεγάλης σημασίας τους για το ρεαλισμό του συνθετικού βίντεο. Για τον ίδιο λόγο χρησιμοποιείται μια τρισδιάστατη αναπαράσταση του ανθρώπινου προσώπου, ώστε να επιτευχθεί η καλύτερη δυνατή εκφραστικότητα. Επιπλέον, με το σχεδιασμό του μοντέλου ώστε να παράγει ταυτόχρονα την ομιλία και το δυναμικό τρισδιάστατο πρόσωπο, το οποίο οδηγεί ένα νευρωνικό μοντέλο renderer, έχουμε εκ κατασκευής συγχρονισμένη ηχητική και οπτική ροή. Συνεπώς, το σύστημά μας διαχωρίζει τη σύνθεση φωτορεαλιστικού ομιλούντος προσώπου με είσοδο κείμενο σε δύο υποεργασίες: οπτικοακουστική παραγωγή ομιλίας και 3Δ προσώπου, και απεικόνιση του 3Δ βίντεο προσώπου σε έγχρωμο, πλήρες βίντεο. Κάθε υποεργασία εκτελείται από μία μονάδα του συστήματός μας. Οι μονάδες αυτές, οπτικοακουστική και φωτορεαλιστική, εκπαιδεύονται ξεχωριστά στα ίδια δεδομένα, και λειτουργούν εν σειρά κατά τη διαδικασία σύνθεσης.

Προεπεξεργασία Δεδομένων

Τα σύνολα δεδομένων που χρησιμοποιούνται για σύνθεση ομιλούντων προσώπων αποτελούνται από βίντεο ατόμων που εκφέρουν προτάσεις, μαζί με τις απομαγνητοφωνήσεις τους. Προκειμένου να εκπαιδεύσουμε το σύστημά μας, εκτελούμε μια προεπεξεργασία στα δεδομένα, ώστε να εξάγουμε απαραίτητες πληροφορίες όπως η ευθυγράμμιση κειμένου-φωνής και η 3Δ ανακατασκευή του προσώπου.

Αρχικά επεξεργαζόμαστε τα καρέ του βίντεο. Χρησιμοποιούμε το μοντέλο MTCNN (Zhang et al. 2016) για να ανιχνεύσουμε τη θέση του προσώπου σε κάθε καρέ και να περικόψουμε την εικόνα σε ανάλυση 256×256 γύρω από το πρόσωπο του ατόμου. Επιπλέον, χρησιμοποιούμε το συνελικτικό μοντέλο FSGAN (Nirkin et al. 2019) για να εξάγουμε τη μάσκα του εσωτερικού του προσώπου. Ο σκοπός είναι να εκπαιδεύσουμε ένα μοντέλο παραγωγής εικόνων μόνο του εσωτερικού του προσώπου, ώστε η μέθοδός μας να είναι ανεξάρτητη από αλλαγές στο υπόβαθρο του καρέ. Πραγματοποιούμε 3Δ ανακατασκευή του κεφαλιού του ατόμου αξιοποιώντας τη μέθοδο SPECTRE (Filntisis et al. 2023), η οποία έχει εκπαιδευτεί με στόχο την ακριβή αποτύπωση κινήσεων του στόματος, χρησιμοποιώντας χαρακτηριστικά ανάγνωσης χειλιών. Έτσι, μπορούμε να κατασκευάσουμε τις εικόνες του 3Δ σχήματος, και την εικόνα Normalized Mean Face Coordinate (NMFC), μια σημασιολογική αναπαράσταση η οποία κωδικοποιεί χρωματικά τη θέση των σημείων του 3Δ πλέγ-



Σχήμα 8: Διαδικασία προεπεξεργασίας των βίντεο του συνόλου δεδομένων, με παράδειγμα ενός καρέ. Το πρόσωπο ανιχνεύεται και περικόπτεται σε ένα τετραγωνικό πλαίσιο. Στη συνέχεια, δημιουργείται μια μάσκα για να απομονωθεί το εσωτερικό του προσώπου. Το κεφάλι ανακατασκευάζεται σε 3Δ, ενώ επίσης προβλέπονται 2Δ σημεία αναφοράς του προσώπου, που περιλαμβάνουν τις άκρες των ματιών και του στόματος. Τέλος, υπολογίζεται ο βέλτιστος μετασχηματισμός που αντιστοιχίζει τα σημεία αναφοράς με ένα πρότυπο. Αυτός ο μετασχηματισμός εφαρμόζεται σε όλες τις παραγόμενες εικόνες.

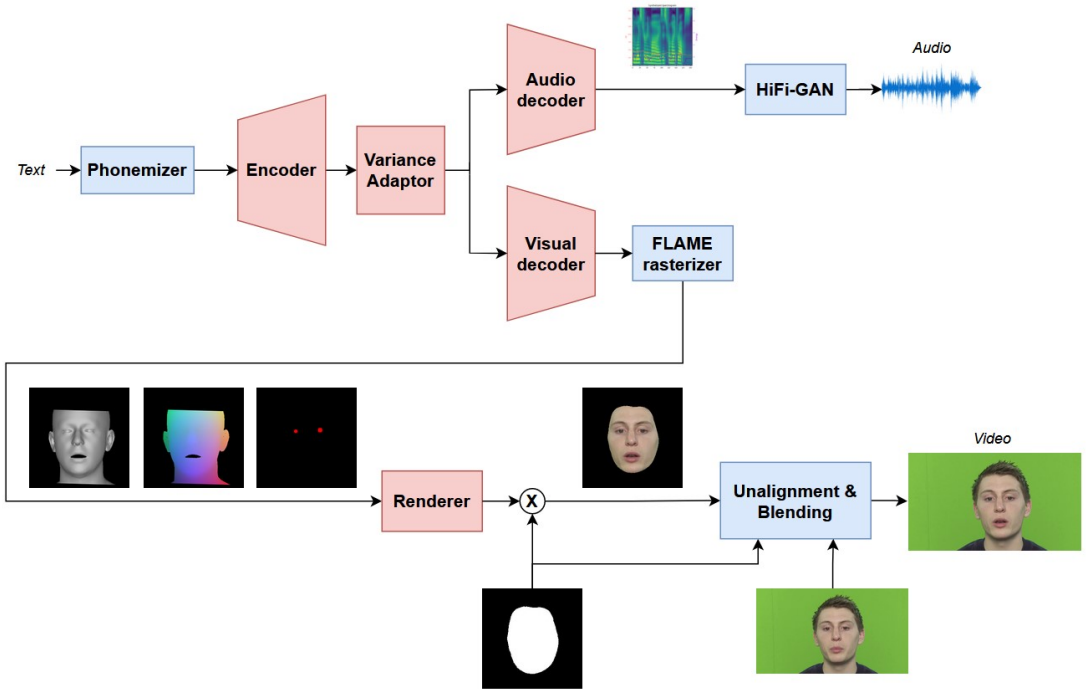
ματος στο χώρο. Τέλος, ανιχνεύουμε τα σημεία αναφοράς του προσώπου με τη χρήση του μοντέλου FAN (Bulat and Tzimiropoulos 2017), και τα χρησιμοποιούμε για να εκτιμήσουμε τον βέλτιστο μετασχηματισμό απεικόνισής τους σε μια πρότυπη γεωμετρία. Με αυτό τον τρόπο μπορούμε να μετασχηματίσουμε όλες τις παραγόμενες εικόνες ώστε το πρόσωπο να είναι ακριβώς στο κέντρο της εικόνας, και να έχει κατά τον δυνατόν πανομοιότυπο μέγεθος και προσανατολισμό. Αυτός ο μετασχηματισμός αποσκοπεί στο να βελτιώσει τη δυνατότητα γενίκευσης του renderer. Η προεπεξεργασία των καρέ απεικονίζεται στο Σχήμα 8.

Στη συνέχεια, επεξεργαζόμαστε τον ήχο του βίντεο. Χρησιμοποιούμε το μοντέλο MFA (McAuliffe et al. 2017) για την ευθυγράμμιση της κυματομορφής ήχου με τη φωνητική ακολουθία του κειμένου. Επιπλέον, υπολογίζουμε τη θεμελιώδη συχνότητα του σήματος ανά ακουστικό πλαίσιο, το φασματογράφημα, καθώς και την ενέργειά του.

Οπτικοακουστική Μονάδα

Η οπτικοακουστική μονάδα επεκτείνει την αρχιτεκτονική ενός μοντέλου σύνθεσης φωνής, προκειμένου να ενσωματώσει και την σύνθεση οπτικής πληροφορίας. Συγκεκριμένα, ακολουθούμε το FastSpeech 2 (Ren et al. 2020), το οποίο χρησιμοποιεί μετασχηματιστές για να μετατρέψει φωνήματα σε χαρακτηριστικά mel. Επεκτείνουμε αυτήν την αρχιτεκτονική πρόβλεψης φασματογραμμάτων προβλέποντας ένα διάνυσμα συντελεστών 3ΔΜΜ ανά καρέ ήχου, παράγοντας συγχρονισμένα ηχητικά και οπτικά χαρακτηριστικά. Η αρχιτεκτονική του συστήματος περιλαμβάνει έναν δεύτερο αποκωδικοποιητή για την παραγωγή των συντελεστών FLAME 3ΔΜΜ, ο οποίος λαμβάνει, μαζί με τον ηχητικό αποκωδικοποιητή, την έξοδο του κωδικοποιητή κειμένου. Αυτή η προσέγγιση αποφεύγει την περιττή επανάληψη, την πιθανή αναντιστοιχία συνόλων δεδομένων, και τη συσσώρευση σφαλμάτων που μπορεί να υπάρχουν σε εν σειρά προσεγγίσεις. Χρησιμοποιούμε το SPECTRE για την 3Δ ανακατασκευή του προσώπου, και μοντελοποιώντας την έκφραση και τις κινήσεις του προσώπου κατά τη διάρκεια της ομιλίας. Η οπτικοακουστική μονάδα περιλαμβάνει τις εξής υπο-μονάδες:

- **Αναλυτής φωνημάτων:** Η υπο-μονάδα αυτή μετατρέπει το κείμενο στην κατάλληλη ακολουθία φωνημάτων, χρησιμοποιώντας το φωνητικό λεξικό του CMU. Για άγνωστες λέξεις, χρησιμοποιεί ένα προεκπαιδευμένο δίκτυο πρόβλεψης των φωνημάτων με βάση την ορθογραφία της λέξης (Park and Kim 2019).
- **Κωδικοποιητής:** Ο κωδικοποιητής λαμβάνει ως είσοδο την ακολουθία φωνημάτων και την μετατρέπει σε μια κρυφή αναπαράσταση μέσω ενός μετασχηματιστή 4 επιπέδων.
- **Προσαρμογέας (Variance Adaptor):** Ο προσαρμογέας εγχέει ακουστική πληροφορία στην κρυφή αναπαράσταση των φωνημάτων, προσθέτοντας διανύσματα που κωδικοποιούν τη θεμελιώδη συχνότητα και την ενέργεια της φωνής. Παράλληλα, χρησιμοποιεί την πληροφορία της διάρκειας κάθε φωνήματος ώστε να επεκτείνει την ακολουθία, η οποία αρχικά είχε μήκος όσα και τα φωνήματα του κειμένου εισόδου, προκειμένου να φτάσει στο κατάλληλο μήκος. Για παράδειγμα, για φωνήματα με μεγαλύτερη διάρκεια, όπως τα φωνήεντα, τα διανύσματα της κρυφής ακολουθίας επαναλαμβάνονται κάποιες φορές, ώστε να ταιριάζουν με τη διάρκεια του φωνήματος.
- **Ηχητικός αποκωδικοποιητής:** Ένας μετασχηματιστής 6 επιπέδων, ακολουθούμενος από ένα γραμμικό επίπεδο. Η λειτουργία του είναι να μετασχηματίζει την κρυφή ακολουθία στο φασματογράφημα της φωνής.
- **Κωδικοποιητής φωνής (Vocoder):** Χρησιμοποιούμε το προ-εκπαιδευμένο HiFi-GAN δίκτυο (Kong et al. 2020a) ώστε να μετατρέψουμε το φασματογράφημα σε κυματομορφή.
- **Οπτικός αποκωδικοποιητής:** Παρόμοια με τον ηχητικό αποκωδικοποιητή, χρησιμοποιείται ένας μετασχηματιστής 4 επιπέδων και ένα γραμμικό επίπεδο για τη μετατροπή των κρυφών χαρακτηριστικών σε χρονοσειρές συντελεστών 3ΔΜΜ. Οι δύο αποκωδικοποιητές επιλέγονται να έχουν διαφορετικά βάθη λόγω της διαφορετικής διαστατικότητας των δεδομένων που πρέπει να μοντελοποιήσουν. Ο πρώτος πρέπει να προβλέψει μια τιμή για 80 συχνοτικές ζώνες, ενώ ο δεύτερος πρέπει να προβλέψει 53 ανεξάρτητες συνιστώσες 3ΔΜΜ.



Σχήμα 9: Η διαδικασία σύνθεσης ομιλούντων προσώπων μέσω του μοντέλου NEUTART. Εκκινώντας από μια φράση κειμένου, η οπτικοακουστική μονάδα (άνω κομμάτι) τη μετατρέπει σε φωνητική ακολουθία, η οποία μετατρέπεται ταυτόχρονα σε φασματογράφημα και 3Δ παραμέτρους προσώπου. Η φωνή συντίθεται από το φασματογράφημα μέσω του HiF-GAN, ενώ οι 3Δ συντελεστές μετατρέπονται σε 3Δ βίντεο μέσω του FLAME. Η φωτορεαλιστική μονάδα (κάτω κομμάτι) χρησιμοποιεί το βίντεο του 3Δ προσώπου για να προβλέψει το έγχρωμο βίντεο, το οποίο θα αντικαταστήσει το πρόσωπο από κάποιο βίντεο αναφοράς. Για απλότητα, απεικονίζουμε μονάχα ένα καρέ. Τα νευρωνικά δίκτυα που βελτιστοποιούμε απεικονίζονται με ροζ χρώμα, ενώ προεκπαιδευμένα δίκτυα απεικονίζονται με γαλάζιο χρώμα.

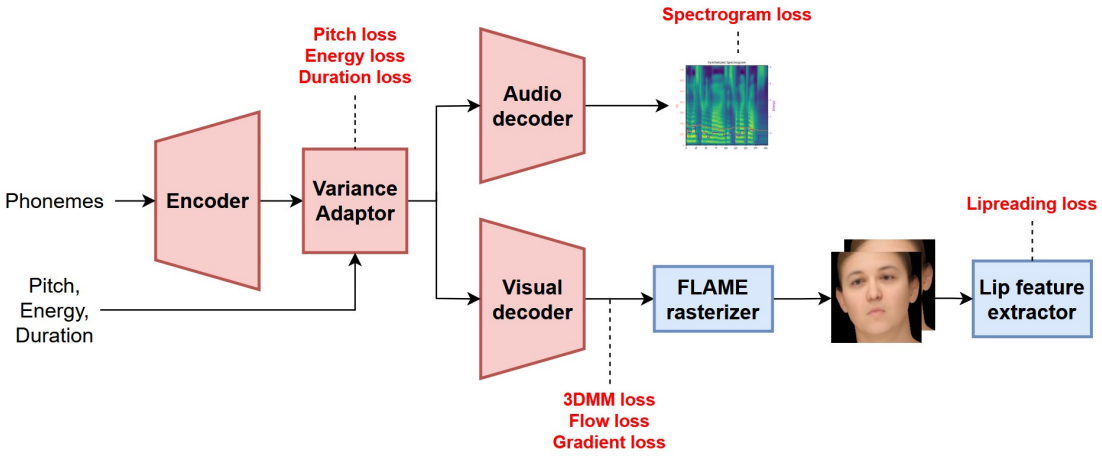
- **FLAME rasterizer:** Χρησιμοποιούμε τον αποκωδικοποιητή και rasterizer του μοντέλου FLAME ώστε να μετατρέψουμε τους συντελεστές 3ΔΜΜ σε 3Δ πλέγμα, και έπειτα να το απεικονίσουμε.

Η αρχιτεκτονική της οπτικοακουστικής μονάδας απεικονίζεται στο άνω μισό του Σχήματος 9. Προκειμένου να την εκπαιδεύσουμε, χρησιμοποιούμε από κάθε πρόταση του συνόλου δεδομένων:

- Την ακολουθία φωνημάτων $p_{1:L}$.
- Τη μέση θεμελιώδη συχνότητα ανά φώνημα $\hat{f}_{1:L}$.
- Τη μέση ενέργεια ανά φώνημα $\hat{e}_{1:L}$.
- Τη διάρκεια κάθε φωνήματος $\hat{d}_{1:L}$.
- Το φασματογράφημα $\hat{\mathbf{F}}$.
- Τους συντελεστές 3ΔΜΜ $\hat{\vec{x}}_n$.
- Το αρχικό βίντεο περικομμένο γύρω από το στόμα του ατόμου $\hat{\mathbf{I}}^M_{1:N}$.
- Τις παραμέτρους κάμερας $\vec{c}_{1:N}$.

Βελτιστοποιούμε τη συνάρτηση κόστους $\mathcal{L}_{av}$, η οποία αποτελείται από:

- **Σφάλματα συχνότητας, ενέργειας, και διάρκειας:** Ακολουθώντας το FastSpeech 2, βελτιστοποιούμε τις προβλέψεις του προσαρμογέα χρησιμοποιώντας μέσο τετραγ-



Σχήμα 10: Η διαδικασία εκπαίδευσης της οπτικοακουστικής μονάδας απαιτεί την ακολουθία φωνημάτων μιας πρότασης, καθώς και τη μέση θεμελιώδη συχνότητα, ενέργεια, και διάρκεια ανά φώνημα. Ο προσαρμογέας εκπαιδεύεται να προβλέπει τις προαναφερθείσες ακουστικές ποσότητες χρησιμοποιώντας τις πραγματικές τιμές για την εκπαίδευση, και τις προβλεπόμενες τιμές κατά την πρόβλεψη. Η έξοδος του ηχητικού αποκωδικοποιητή είναι το φασματογράφημα mel, το οποίο χρησιμοποιείται για τον υπολογισμό του σχετικού σφάλματος. Επίσης, η έξοδος του οπτικού αποκωδικοποιητή είναι η ακολουθία 3ΔMM, η οποία απεικονίζεται σε ένα 3D πλέγμα χρησιμοποιώντας το FLAME, και στη συνέχεια προβάλλεται σε καρέ βίντεο, από τα οποία μπορούν να εξαχθούν χαρακτηριστικά ανάγνωσης χειλιών. Επισημαίνουμε ότι υπάρχει τόσο ακουστική όσο και οπτική επίβλεψη, μοντελοποιώντας την οπτικοακουστική σύνθεση φωνής ως πολυτροπική διαδικασία.

ωνικό σφάλμα στις ποσότητες που καλείται να προβλέψει. Για παράδειγμα, το σφάλμα συχνότητας υπολογίζεται:

$$\mathcal{L}_{pitch} = \mathbb{E}_l[||\hat{f}_l - f_l||_2^2] \tag{5}$$

με f_l την προβλεπόμενη συχνότητα.

- **Σφάλμα φασματογραφήματος:** Χρησιμοποιούμε το μέσο απόλυτο σφάλμα για το φασματογράφημα:

$$\mathcal{L}_{mel} = ||\hat{\mathbf{F}} - \mathbf{F}||_1 \tag{6}$$

- **Σφάλμα 3ΔMM:** Μέσο τετραγωνικό σάλμα πρόβλεψης των συντελεστών 3ΔMM.
- **Σφάλμα κλίσης:** Ποινικοποιούμε τις απότομες αλλαγές στις χρονοσειρές των 3ΔMM συντελεστών, για να εξασφαλίσουμε πιο ομαλό αποτέλεσμα πρόβλεψης, με τον όρο:

$$\mathcal{L}_{grad} = \mathbb{E}_n[||\vec{x}_{n+1} - \vec{x}_n||_2^2] \tag{7}$$

- **Σφάλμα ροής:** Χρησιμοποιούμε το σφάλμα χρονικών διαφορών βασιζόμενοι στους Hussen Abdelaziz et al. (2020):

$$\mathcal{L}_{flow} = \mathbb{E}_n[||(\hat{\vec{x}}_{n+1} - \hat{\vec{x}}_n) - (\vec{x}_{n+1} - \vec{x}_n)||_2^2] \tag{8}$$

όπου $\hat{\vec{x}}_n$ είναι η πραγματική τιμή του διανύσματος τιμών FLAME, και $\vec{x}_n$ είναι η προβλεπόμενη τιμή για το καρέ n.

- **Σφάλμα ανάγνωσης χειλιών:** Ακολουθώντας την εκπαίδευση του SPECTRE, χρησιμοποιούμε ένα μοντέλο ανάγνωσης χειλιών (Ma et al. 2022) για να εξάγουμε χαρακτηριστικά στο 3Δ βίντεο, και υπολογίζουμε την απόστασή τους από τα αντίστοιχα χαρακτηριστικά του

πραγματικού βίντεο (μετά από περικοπή γύρω από το στόμα). Ως μέτρο απόστασης χρησιμοποιούμε την απόσταση συνημιτόνου, καθώς είναι καταλληλότερη από την Ευκλείδια απόσταση για τον υπολογισμό σφαλμάτων σε υψηλής διάστασης χώρους. Υπολογίζουμε:

$$\mathcal{L}_{lip} = \mathbb{E}_n \left[1 - \frac{\hat{\vec{f}}_n \cdot \vec{f}_n}{||\hat{\vec{f}}_n|| \; ||\vec{f}_n||} \right] \tag{9}$$

όπου $\hat{\vec{f}}_n,\ \vec{f}_n \in \mathbb{R}^{512}$ είναι τα χαρακτηριστικά ανάγνωσης χειλιών στο καρέ n, για το πραγματικό και το προβλεπόμενο 3Δ βίντεο, αντίστοιχα.

- **Σφάλμα κανονικοποίησης εκφράσεων:** Όπως παρατηρούν οι Filntisis et al. (2023), το σφάλμα ανάγνωσης χειλιών μπορεί να προκαλέσει ταλάντωση στις χρονοσειρές των εκφράσεων. Επομένως, χρησιμοποιούμε το παρακάτω σφάλμα κανονικοποίησης για να ποινικοποιήσουμε το μέτρο των προβλεπόμενων συντελεστών έκφρασης $\vec{\psi}_n$:

$$\mathcal{L}_{reg} = 10^{-3} \mathbb{E}_n [w_n ||\vec{\psi}_n||_2^2] \tag{10}$$

χρησιμοποιώντας τα εμπειρικά καθορισμένα βάρη:

$$w_n = \begin{cases} 1, & ||\vec{\psi}_n||_2^2 < 40 \\ 2, & ||\vec{\psi}_n||_2^2 > 40 \end{cases} \tag{11}$$

Φωτορεαλιστική Μονάδα

Το βασικό στοιχείο αυτής της μονάδας είναι ένας GAN renderer $\mathcal{R}$, ο οποίος εκπαιδεύεται να προβλέπει εικόνες προσώπου χρησιμοποιώντας τις 3Δ απεικονίσεις τους ως είσοδο. Ακολουθούμε την αρχιτεκτονική του Head2Head++ (Doukas et al. 2021b) που προσαρμόστηκε για τις ανάγκες του Neural Emotion Director (Paraperas Papantoniou et al. 2022), η οποία μας επιτρέπει να τροποποιήσουμε το οπτικό περιεχόμενο ενός βίντεο αναφοράς με ένα ομιλούν άτομο. Ο renderer υλοποιείται με μια συνελικτική αρχιτεκτονική που επιτελεί μια εργασία μετάφρασης από εικόνα σε εικόνα, βασισμένη σε GANs.

Τυπικά, ο renderer προβλέπει ένα καρέ βίντεο $\mathbf{I}_n \in \mathbb{R}^{W \times H \times 3}$, με βάση την εικόνα $\mathbf{S}_n \in \mathbb{R}^{W \times H \times 3}$ των 3D σχημάτων, την εικόνα NMFC $\mathbf{N}_n \in \mathbb{R}^{W \times H \times 3}$, την εικόνα των ματιών $\mathbf{E}_n \in \mathbb{R}^{W \times H \times 3}$, καθώς και τα δύο προηγούμενα καρέ. Για απλότητα, θα σημειώσουμε τη συνένωση καναλιών των $(\mathbf{S}_n, \mathbf{N}_n, \mathbf{E}_n)$ ως $\mathbf{X}_n \in \mathbb{R}^{W \times H \times 9}$.

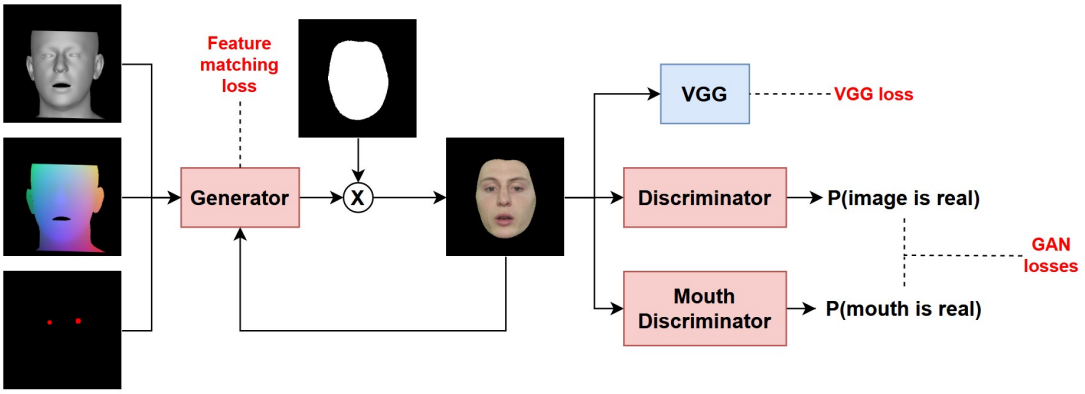
Ως αρχικές συνθήκες, χρησιμοποιούμε $\mathbf{I}_{-2} = \mathbf{I}_{-1} = \mathbf{I}_0$. Η συνένωση της ακολουθίας των παραγόμενων καρέ κατά μήκος της διάστασης του χρόνου οδηγεί στο βίντεο του προσώπου $\mathbf{I}_{1:N}$, το οποίο αναμιγνύεται με κάποιο βίντεο αναφοράς για να συμπληρωθεί το φόντο.

Η διαδικασία εκπαίδευσης με το ανταγωνιστικό σχήμα GAN αποτελείται από τον προαναφερθέντα γεννήτορα $\mathcal{R}$, καθώς και από έναν εκτιμητή εικόνας $\mathcal{D}$ και έναν ειδικό εκτιμητή στόματος $\mathcal{D}_M$, με στόχο τη βελτίωση του ρεαλισμού στην περιοχή του στόματος. Ακολουθώντας την υλοποίηση των Paraperas Papantoniou et al. (2022), ο γεννήτορας κατασκευάζεται με παρόμοιο τρόπο με το Vid2Vid (Wang et al. 2018a), ενώ οι εκτιμητές υιοθετούν τις αρχιτεκτονικές τους από το Pix2PixHD (Wang et al. 2018b).

Η φωτορεαλιστική μονάδα εκπαιδεύεται με τις παρακάτω συναρτήσεις σφάλματος.

- **Ανταγωνιστικό σφάλμα:** Για την ανταγωνιστική εκπαίδευση, χρησιμοποιούμε την απώλεια του Least Squares GAN που προτείνεται από τον Mao et al. (2017), η οποία βελτιώνει τη σταθερότητα της εκπαίδευσης. Ο στόχος του γεννήτορα $\mathcal{R}$ είναι να κάνει τον εκτιμητή να προβλέπει με υψηλή πιθανότητα ότι ένα πλαστό δείγμα είναι πραγματικό:

$$\mathcal{L}_{GAN}^{\mathcal{R}} = \frac{1}{2} \mathbb{E}_n [(\mathcal{D}(\mathbf{X}_n, \mathbf{I}_n) - 1)^2 + (\mathcal{D}_M(\mathbf{X}_n^M, \mathbf{I}_n^M) - 1)^2] \tag{12}$$



Σχήμα 11: Ένα στιγμιότυπο της διαδικασίας εκπαίδευσης της φωτορεαλιστικής μονάδας για ένα καρέ. Η εκπαίδευση ακολουθεί το ανταγωνιστικό πρωτόκολλο των GANs, χρησιμοποιώντας έναν γεννήτορα (renderer $\mathcal{R}$) και δύο εκτιμητές. Οι είσοδοι του renderer είναι το ευθυγραμμισμένο σχήμα, NMFC, και η εικόνα ματιών, καθώς και τα πρόσωπα από τα δύο προηγούμενα καρέ, που απεικονίζονται με έναν απλό βρόχο ανάδρασης. Στην έξοδο του renderer εφαρμόζεται η μάσκα, και το εναπομένον εσωτερικό του προσώπου λειτουργεί ως είσοδος στους δύο εκτιμητές.

όπου $\mathbf{I}_n$ είναι η τρέχουσα έξοδος του γεννήτορα, και $\mathbf{I}_n^M$ είναι η ίδια εικόνα, περικομμένη γύρω από το στόμα. Αντίθετα, κάθε εκτιμητής πρέπει να προβλέπει χαμηλή πιθανότητα για πλαστά δεδομένα και υψηλή για πραγματικά δεδομένα. Για παράδειγμα, ο εκτιμητής που βλέπει το σύνολο του συνθετικού προσώπου καλείται να βελτιστοποιήσει:

$$\mathcal{L}_{GAN}^{\mathcal{D}} = \frac{1}{2}\mathbb{E}_n[(D(\mathbf{X}_n, \hat{\mathbf{I}}_n) - 1)^2 + \mathcal{D}(\mathbf{X}_n, \mathbf{I}_n)^2] \tag{13}$$

- **Σφάλμα VGG:** Αυτή η αντιληπτική συνάρτηση σφάλματος χρησιμοποιεί το δίκτυο Visual Geometry Group (Simonyan and Zisserman 2015) για να εξάγει οπτικά χαρακτηριστικά και να βρει την απόσταση μεταξύ τους. Μπορεί να εκφραστεί ως:

$$\mathcal{L}_{VGG} = \mathbb{E}_n \left[\sum_i \frac{1}{M_i} ||\mathcal{F}_i(\hat{\mathbf{I}}_n) - \mathcal{F}_i(\mathbf{I}_n)||_1 \right] \tag{14}$$

όπου $\mathcal{F}_i$ είναι το i-οστό επίπεδο του δικτύου VGG.

- **Σφάλμα ταιριάσματος χαρακτηριστικών:** Ακόμα ένα αντιληπτικό σφάλμα που ποινικοποιεί τις διαφορές μεταξύ χαρακτηριστικών για πραγματικές και παραγόμενες εικόνες. Μπορεί να εκφραστεί με παρόμοιο τρόπο με το σφάλμα VGG, αλλά χρησιμοποιεί τις εξόδους των στρωμάτων $\mathcal{D}_i$ του εκτιμητή:

$$\mathcal{L}_{FM} = \mathbb{E}_n \left[\sum_i \frac{1}{M_i} ||\mathcal{D}_i(\hat{\mathbf{I}}_n) - \mathcal{D}_i(\mathbf{I}_n)||_1 \right] \tag{15}$$

Συνολικά, η συνάρτηση σφάλματος που χρησιμοποιούμε για την εκπαίδευση του renderer $\mathcal{R}$ είναι:

$$\mathcal{L}_{ph} = \mathcal{L}_{GAN}^{\mathcal{R}} + 10\mathcal{L}_{VGG} + 10\mathcal{L}_{FM} \tag{16}$$

Τα βάρη επιλέχθηκαν με βάση τις προαναφερθείσες προηγούμενες εργασίες. Η συνολική διαδικασία εκπαίδευσης της φωτορεαλιστικής μονάδας παρουσιάζεται στο Σχήμα 11.

6 Πειράματα

Σύνολα Δεδομένων

Καθώς αποσκοπούμε στην ανάπτυξη ενός μοντέλου οπτικοακουστικής σύνθεσης ομιλούντων προσώπων από κείμενο, χρειαζόμαστε κατάλληλα σύνολα δεδομένων, τα οποία αφενός έχουν καλή ποιότητα ήχου, αφετέρου περιλαμβάνουν και τις απομαγνητοφωνήσεις των δειγμάτων βίντεο. Το σύνολο οπτικοακουστικών δεδομένων TCD-TIMIT (Harte and Gillen 2015), το οποίο αποτελείται από υψηλής ποιότητας δείγματα 62 ομιλητών γυρισμένα σε συνθήκες εργαστηρίου, είναι το πλέον κατάλληλο για την εκπαίδευση αυτού του μοντέλου.

Ωστόσο, χρησιμοποιούμε και το σύνολο ηχητικών δεδομένων LJSpeech (Ito and Johnson 2017), το οποίο αποτελείται από ηχογραφήσεις μιας ομιλήτριας που διαβάζει αποσπάσματα από βιβλία. Αρχικά, εκπαιδεύουμε ένα TTS σύστημα που έχει ίδια αρχιτεκτονική με την οπτικοακουστική μονάδα, αλλά χωρίς τον οπτικό αποκωδικοποιητή. Με αυτό τον τρόπο, μπορούμε να αρχικοποιήσουμε τα βάρη των υπόλοιπων υπο-μονάδων στα βάρη που βελτιστοποιήθηκαν για τη σύνθεση φωνής στο LJSpeech, εκτελώντας μεταφορά μάθησης (Yosinski et al. 2014). Έτσι, η μετέπειτα εκπαίδευση στο TCD-TIMIT επιτυγχάνει καλύτερη σύγκλιση. Η φωτορεαλιστική μονάδα εκπαιδεύεται χρησιμοποιώντας τα καρέ από τα ίδια βίντεο του TCD-TIMIT.

Επιπλέον, εκτελούνται πειράματα στο μικρό σύνολο δεδομένων LIPS2008 (Theobald et al. 2008), καθώς και σε κάποια βίντεο "in-the-wild", δηλαδή με φυσικές συνθήκες κινηματογράφησης, όπως κάποια δείγματα από το σύνολο δεδομένων HDTF (Zhang et al. 2021) και μερικά δημόσια βίντεο από το YouTube.

Αξιολόγηση

Αρχικά εκπαιδεύσαμε την οπτικοακουστική μονάδα με πολλούς ομιλητές από το TCD-TIMIT, ακολουθώντας την πολυομιλητική αρχιτεκτονική του FastSpeech 2, προκειμένου να επωφεληθούμε από την πληθώρα δεδομένων ολόκληρου του συνόλου. Η διάρκεια υλικού ανά ομιλητή είναι λιγότερη από 10 λεπτά, το οποίο είναι πολύ λίγο για την εκπαίδευση ενός μοντέλου σύνθεσης ομιλίας από την αρχή. Επιπλέον, για να βελτιώσουμε περαιτέρω την ποιότητα του παραγόμενου ήχου, αρχικοποιήσαμε τον κωδικοποιητή και τον ηχητικό αποκωδικοποιητή με τα βάρη ενός όμοιου TTS μοντέλου εκπαιδευμένο στο LJSpeech. Ο οπτικός αποκωδικοποιητής αρχικοποιήθηκε τυχαία, και όλα τα μοντέλα εκπαιδεύτηκαν για 50.000 επαναλήψεις σε κάθε πείραμα. Η οπτικοακουστική μονάδα μπορεί να χρησιμοποιηθεί όπως είναι, ή να γίνει αναπροσαρμογή σε έναν συγκεκριμένο ομιλητή για λίγες επαναλήψεις, προκειμένου να δημιουργήσουμε προσωποποιημένα μοντέλα.

Η αξιολόγηση των ομιλούντων προσώπων αποτελεί πρόκληση, καθώς οι στατιστικές μετρικές σφάλματος δε συνάδουν πάντα με τις ανθρώπινες αξιολογήσεις (Chen et al. 2020a). Επομένως, αξιολογούμε αντικειμενικά το μοντέλο μας χρησιμοποιώντας τόσο στατιστικές όσο και αντιληπτικές μετρικές, και επιπλέον αξιολογούμε υποκειμενικά με μια μελέτη χρηστών.

Θέλουμε να αξιολογήσουμε το μοντέλο μας βασιζόμενο στην ηχητική και φωτορεαλιστική έξοδό του, γι' αυτό συγκρίνουμε με πρόσφατες μεθόδους που παράγουν RGB βίντεο. Ως αποτέλεσμα, δεν συγκρίνουμε το μοντέλο μας με το AVTacotron2 (Abdelaziz et al. 2021) ή το UniFLG (Mitsui et al. 2023), τα οποία εκτελούν οπτικοακουστική σύνθεση, επειδή η έξοδός τους είναι μια μη φωτορεαλιστική 3Δ απεικόνιση. Επιπλέον, και τα δύο είναι ιδιόκτητες υλοποιήσεις, ενώ εμείς θα θέλαμε να πειραματιστούμε με ανοικτού κώδικα έργα, ώστε να υπάρχει δυνατότητα αναπαραγωγής και επιβεβαίωσης των πειραμάτων.

Η έρευνα που διεξάγεται στην παρούσα Διπλωματική Εργασία στοχεύει στη δημιουργία οπτικοακουστικής ομιλίας από κείμενο. Ωστόσο, δεδομένου ότι δεν υπάρχουν πρόσφατες δημόσια διαθέσιμες μέθοδοι που να λειτουργούν με βάση το κείμενο, η δίκαιη επιλογή θα ήταν να συγκρίνουμε με μοντέλα που βασίζονται στον ήχο και παράγουν φωτορεαλιστικό βίντεο, αλλά χρησιμοποιώντας συνθετικό ήχο για να τα δειγματοληπτήσουμε. Επιλέξαμε να συγκρίνουμε το μοντέλο μας (NT) με τα ακόλουθα δημοφιλή μη προσωποποιημένα (few-shot) μοντέλα συγχρονισμού χειλιών.

- Wav2Lip (W2L), των Prajwal et al. (2020)

- SadTalker (ST), από τους Zhang et al. (2022c)
- VideoReTalking (VRT), των Cheng et al. (2022)

Η δειγματοληψία από αυτά τα μοντέλα πραγματοποιήθηκε χρησιμοποιώντας ήχο από το FastSpeech 2, με την ίδια αρχιτεκτονική κωδικοποιητή και ηχητικού αποκωδικοποιητή όπως στην οπτικοακουστική μας μονάδα. Το γεγονός ότι οι παραπάνω μέθοδοι δεν είναι προσωποποιημένες σημαίνει πως έχουν είναι σχεδιασμένες να έχουν καλύτερη ικανότητα γενίκευσης, ωστόσο ενδεχομένως να μην είναι τα πλέον κατάλληλα για σύνθεση με υψηλό ρεαλισμό.

Αντικειμενική Αξιολόγηση

Η αντικειμενική αξιολόγηση διενεργείται σε τρία τυχαία επιλεγμένα άτομα από το TCD-TIMIT, με δείγματα που εξάγονται από προσωποποιημένες εκδοχές του μοντέλου μας. Διάφορες μετρικές χρησιμοποιούνται, συμπεριλαμβανομένης της μέσης απόστασης mel cepstrum (MCD), του ποσοστού σφάλματος χαρακτήρων μέσω αναγνώρισης φωνής (ACER), του σφάλματος στα σημεία ενδιαφέροντος γύρω από το στόμα καθώς και την ταχύτητάς τους (LMD και LLVE). Επιπλέον αντιληπτικά σφάλματα εξάγονται μέσω ανάγνωσης χειλιών με το μοντέλο AV-HuBERT (Shi et al. 2022a; Shi et al. 2022b), το οποίο μας επιτρέπει να εξάγουμε το ποσοστό σφάλματος χαρακτήρων (VCER) και το ποσοστό σφάλματος οπτικών φωνημάτων (VER). Ο φωτορεαλισμός αξιολογείται χρησιμοποιώντας την απόσταση Fréchet (FID). Οι συγκρινόμενες ακολουθίες ευθυγραμμίζονται χρησιμοποιώντας τη δυναμική στρέβλωση χρόνου για να αφαιρεθεί η ασυμφωνία που προκαλείται από τις τοπικές χρονικές μετατοπίσεις. Τα αποτελέσματα της αξιολόγησης παρουσιάζονται στον Πίνακα 1, με όλες τις μετρικές να υποδεικνύουν καλύτερη απόδοση όταν η τιμή τους είναι χαμηλότερη.

Τα αποτελέσματα υποδεικνύουν ότι το NEUTART μπορεί να παράγει ρεαλιστικά ομιλούντα πρόσωπα με πολύ καλή άρθρωση λόγου, τόσο ακουστικά όσο και οπτικά. Συγκρίνουμε επίσης την απόδοση του NEUTART στη σύνθεση ήχου σε σύγκριση με το αντίστοιχο σύστημα TTS, δηλαδή το μοντέλο FastSpeech 2 που χρησιμοποιήσαμε για να δειγματοληπτήσουμε τις μεθόδους συγχρονισμού χειλιών. Παρουσιάζουμε τις μετρικές καθαρής σύνθεσης φωνής στον Πίνακα 2. Τα αποτελέσματα υποδεικνύουν ότι η συμπερίληψη οπτικής επίβλεψης μπορεί να βελτιώσει την ποιότητα του παραγόμενου ήχου, ειδικά όσον αφορά την κατανόηση του λόγου, υποστηρίζοντας έτσι την αποτελεσματικότητα της ταυτόχρονης μάθησης.

Μελέτη Χρηστών

Πραγματοποιήσαμε επίσης μια μελέτη χρηστών, συγκρίνοντας τη μέθοδό μας με το FastSpeech 2 ως προς τον ρεαλισμό του ήχου, καθώς και των προαναφερθέντων μεθόδων προς τον οπτικοακουστικό ρεαλισμό. Δημιουργήσαμε ένα σύνολο από φωνητικά πλούσιες προτάσεις που χρησιμοποιήθηκαν για τη δημιουργία των δειγμάτων από κάθε μέθοδο, χρησιμοποιώντας δύο τυχαία επιλεγμένους ομιλητές.

Με αυτά τα δείγματα κατασκευάσαμε ένα ερωτηματολόγιο με ζευγάρια προτίμησης. Για κάθε ερώτηση που βασίζεται στον ήχο, παρουσιάζονταν δύο ηχητικά αρχεία και ζητούνταν από τον χρήστη να επιλέξουν αυτό που ακούγεται πιο ρεαλιστικό. Για το οπτικοακουστικό μέρος, παρουσιάζονταν δύο συνθετικά βίντεο (ένα από τη μέθοδό μας, και ένα από κάποια εκ των άλλων) και οι χρήστες κλήθηκαν ξανά να επιλέξουν αυτό που θεωρούν πιο ρεαλιστικό. Σημειώνουμε ότι σε κάθε ερώτηση παρείχαμε στους χρήστες την απομαγνητοφώνηση της πρότασης, καθώς και μια εικόνα του ομιλητή όταν επρόκειτο για αξιολόγηση βίντεο. Η επιλογή της άλλης συγκρινόμενης μεθόδου και η σειρά των βίντεο στο ζεύγος ήταν τυχαίες σε κάθε ερώτηση. Κάθε χρήστης απάντησε συνολικά σε 4 ερωτήσεις με βάση τον ήχο και 15 ερωτήσεις με βάση το βίντεο. Συνολικά, 21 χρήστες συμπλήρωσαν το ερωτηματολόγιο και τα αποτελέσματα φαίνονται στον Πίνακα 3.

Παρατηρούμε ότι η μέθοδός μας αξιολογείται σταθερά ως η πιο ρεαλιστική. Το SadTalker αξιολογείται ως το δεύτερο καλύτερο μοντέλο, και στη συνέχεια ακολουθεί το VideoReTalking. Επίσης, αναδεικνύεται η αποτελεσματικότητα της πολυτροπικής επίβλεψης για τη σύνθεση φωνής, εφόσον το οπτικοακουστικό μοντέλο παράγει ομιλία που προτιμάται έναντι της εξόδου ενός απλού συστή-

ID		MCD (dB)	ACER (%)	LMD	LMVE	FID	VCER (%)	VER (%)
	Gold	-	[6.06]	-	-	-	[84.45]	[75.96]
	W2L	44.21	12.94	1.3125	0.3238	**18.36**	76.78	**68.30**
38F	ST	44.21	12.94	14.3464	0.4238	221.49	79.82	73.33
	VRT	44.21	12.94	1.6474	0.3150	37.32	80.76	75.48
	Ours	**43.41**	**10.94**	**1.1813**	**0.2889**	38.14	**74.70**	68.78
	Gold	-	[29.64]	-	-	-	[88.68]	[78.52]
	W2L	42.58	**25.01**	**1.0609**	**0.2805**	**18.45**	82.21	73.81
42M	ST	42.58	**25.01**	7.0019	0.4010	167.47	80.42	73.09
	VRT	42.58	**25.01**	1.5036	0.2753	25.34	78.97	71.38
	Ours	**42.36**	32.50	1.2073	0.2867	22.91	**78.64**	**71.15**
	Gold	-	[7.36]	-	-	-	[88.53]	[79.42]
	W2L	**43.50**	18.07	1.9305	0.4479	**17.73**	87.62	81.75
49F	ST	**43.50**	18.07	5.3592	0.5820	139.10	83.79	77.41
	VRT	**43.50**	18.07	**1.9215**	0.4347	29.47	84.76	78.15
	Ours	43.64	**16.93**	1.9576	**0.4132**	25.06	**76.88**	**72.02**
	Gold	-	[14.35]	-	-	-	[87.22]	[77.97]
	W2L	43.43	**18.67**	**1.4346**	0.3507	**18.18**	82.20	74.62
Mean	ST	43.43	**18.67**	8.9025	0.4689	176.02	81.34	74.61
	VRT	43.43	**18.67**	1.6908	0.3417	30.71	81.50	75.00
	Ours	**43.14**	20.12	1.449	**0.3296**	28.70	**76.74**	**70.65**

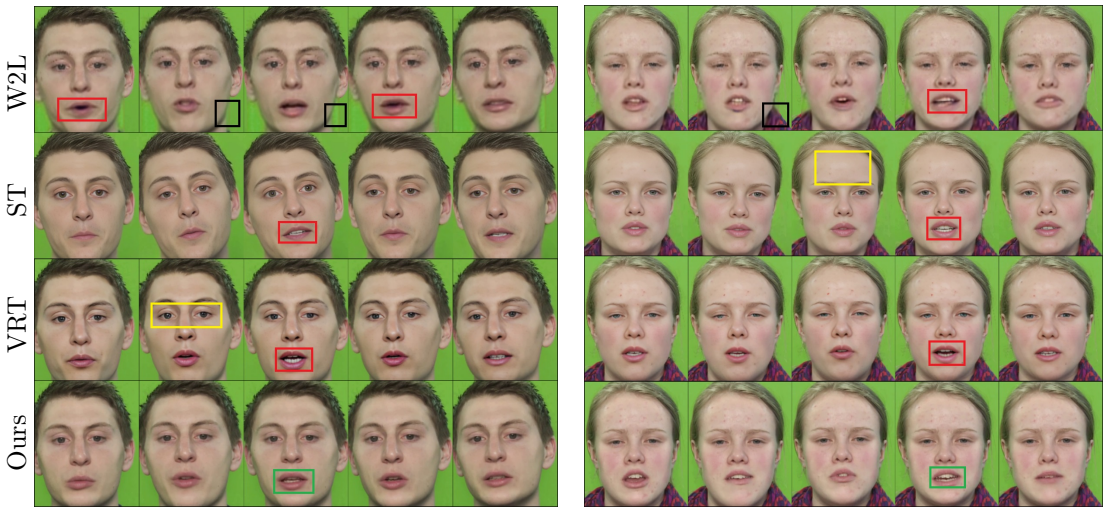
Πίνακας 1: Μετρικές ποιότητας των συνθετικών ήχων και βίντεο σε 3 τυχαία άτομα από το TCD-TIMIT. Η μέθοδός μας παρουσιάζει πολύ καλά αποτελέσματα όσον αφορά τις μετρικές των σημείων ενδιαφέροντος και το FID. Το Wav2Lip μπορεί να έχει χαμηλότερο FID, αλλά παρουσιάζει σημαντικά χειρότερα αποτελέσματα από άλλες μεθόδους όσον αφορά την ανθρώπινη αξιολόγηση, λόγω του ορατού πλαισίου στο στόμα του ατόμου. Τέλος, η μέθοδός μας είναι σταθερά ανώτερη όσον αφορά την ανάγνωση των χειλιών.

Method	MCD (dB)	ACER (%)
Gold	-	[16.21]
FastSpeech 2	**40.13**	27.04
Ours	40.24	**24.85**

Πίνακας 2: Παρατηρούμε ότι η πολυτροπικότητα ευνοεί τη σύνθεση φωνής ως προς την καθαρότητα του λόγου. Η ομιλία που συντίθεται από τη μέθοδό μας, που έχει εκπαιδευτεί με ακουστική και οπτική επίβλεψη, είναι πιο εύκολα κατανοητή σε σχέση με την ομιλία που συντίθεται από όμοια αρχιτεκτονική TTS.

	SadTalker	VideoRetalking	Wav2Lip	FastSpeech 2 (audio only)
NEUTART	**66** / 54	**74** / 46	**87** / 33	**59** / 37
	55.0% / 45.0%	**61.7%** / 38.3%	**72.5%** / 27.5%	**61.5%** / 38.5%

Πίνακας 3: Αποτελέσματα μελέτης χρηστών με σχήμα προτίμησης A/B. Τα αποτελέσματα δείχνουν ότι το NEUTART (αριστερά) προτιμήθηκε A φορές, ενώ η ανταγωνιστική μέθοδος (δεξιά) προτιμήθηκε B φορές, με συνολικά $A + B$ ζευγάρια αξιολόγησης. Παρουσιάζεται επίσης η αντίστοιχη ποσοστιαία αναλογία. Οι χρήστες αξιολόγησαν σταθερά τη μέθοδό μας ως πιο ρεαλιστική από τις ανταγωνιστικές, τόσο όσον αφορά την σύνθεση φωνής όσο και τη πλήρη σύνθεση βίντεο.



Σχήμα 12: Σύγκριση της μεθόδου μας έναντι των προηγούμενων, σε εργαστηριακές συνθήκες. Το Wav2Lip έχει χαμηλή ανάλυση και εμφανίζει ένα ορθογώνιο ψεύδεργο γύρω από το στόμα (επισημασμένο με μαύρο). Το SadTalker και το VideoReTalking παράγουν καρέ με πολύ καλύτερη ανάλυση, αφού χρησιμοποιούν κατάλληλο δίκτυο ενίσχυσης του προσώπου, ωστόσο αυτή η βελτίωση προκαλεί ψεύδεργα που αλλάζουν την ταυτότητα του ατόμου. Για παράδειγμα, η επιδερμίδα της ομιλήτριας είναι υπερβολικά λεία, ενώ το χρώμα των ματιών του ομιλητή αλλάζει (επισημασμένο με κίτρινο). Και οι τρεις μέθοδοι παράγουν κάποια καρέ με μη ρεαλιστικό εσωτερικό του στόματος, ενώ φαίνεται πως αλλάζουν και τα δόντια του ατόμου (επισημασμένο με κόκκινο). Η μέθοδός μας παράγει καρέ με καλοσχηματισμένα χείλη, δόντια, και εσωτερικό του στόματος, χωρίς να δημιουργεί οποιαδήποτε αλλαγή στο πρόσωπο των ατόμων.

ματος σύνθεσης φωνής. Το πρώτο δείγμα από κάθε ομιλητή παρουσιάζεται στον πίνακα 12, όπου σχολιάζουμε ποιοτικά τα οπτικά αποτελέσματα.

Μελέτη Αφαίρεσης

Μελετάμε την επίπτωση κάθε επιπρόσθετου οπτικού σφάλματος πραγματοποιώντας μελέτη αφαίρεσης στο TCD-TIMIT. Αξιολογούμε τα συνθετικά βίντεο της οπτικοακουστικής μονάδας, με τις οπτικές μετρικές να υπολογίζονται από τις εικόνες των τρισδιάστατων ανακατασκευών. Παρουσιάζουμε τα αποτελέσματα στον Πίνακα 4. Οι περισσότερες μετρικές, ιδιαίτερα οι μετρικές ανάγνωσης χειλιών, είναι χαμηλότερες όταν το μοντέλο εκπαιδεύεται χρησιμοποιώντας όλα τα οπτικά σφάλματα.

$\mathcal{L}_{lip}$	$\mathcal{L}_{grad}$	$\mathcal{L}_{flow}$	MCD (dB)	ACER (%)	LMD	LMVE	VCER (%)	VER (%)
✗	✗	✓	41.98	**21.91**	0.5053	0.3502	82.40	77.90
✗	✓	✗	41.91	22.88	0.6856	0.3602	85.66	80.00
✗	✓	✓	**40.31**	25.05	**0.4318**	**0.3261**	77.15	70.99
✓	✓	✓	**40.31**	25.40	0.5063	0.4203	**77.05**	**70.66**

Πίνακας 4: Μελετούμε την αποτελεσματικότητα κάθε επιπλέον όρου που εισάγουμε για οπτική επίβλεψη στην οπτικοακουστική μονάδα. Το σφάλμα ανάγνωσης χειλιών αυξάνει το ρεαλισμό της άρθρωσης, όπως ήταν αναμενόμενο. Επιπλέον, τα σφάλματα κλίσης και ροής βελτιώνουν την πρόβλεψη σημείων ενδιαφέροντος, που σημαίνει ότι αποτυπώνεται πιο πιστά η γεωμετρία του προσώπου. Συνεπώς, χρησιμοποιούμε και τις τρεις συναρτήσεις σφάλματος, μαζί με το απλό σφάλμα πρόβλεψης των 3ΔΜΜ συντελεστών.

	SadTalker	VideoRetalking	Wav2Lip
NEUTART	**135** / 21 **86.54%** / 13.46%	65 / **91** 41.67% / **58.33%**	**107** / 49 **68.59%** / 31.41%

Πίνακας 5: Μελέτη χρηστών σε βίντεο από μη εργαστηριακές συνθήκες. Τα αποτελέσματα δείχνουν ότι το NEUTART (αριστερά) προτιμήθηκε A φορές, ενώ η ανταγωνιστική μέθοδος προτιμήθηκε B φορές, με συνολικά $A + B$ ζευγάρια αξιολόγησης. Ακόμα και υπό φυσικές συνθήκες, η μέθοδός μας παράγει πολύ ενθαρρυντικά αποτελέσματα, εφόσον οι χρήστες προτίμησαν σημαντικά το NEUTART έναντι του Wav2Lip και του SadTalker.

Σχήμα 13: Συγκρίσεις σε μη εργαστηριακές συνθήκες. Οι προηγούμενες μέθοδοι παρουσιάζουν τις ίδιες αδυναμίες, εφόσον δεν επηρεάζονται από τη διαφοροποίηση μεταξύ εργαστηριακών ή μη συνθηκών. Το οπτικό αποτέλεσμα της μεθόδου μας παραμένει πολύ ρεαλιστικό και συγχρονισμένο με τον ήχο. Παρόλα αυτά, η ποιότητα του ήχου δεν είναι εξίσου καλή σε σύγκριση με των δειγμάτων που προέρχονται από εκπαίδευση σε εργαστηριακές συνθήκες.

Πειράματα σε Μη Εργαστηριακές Συνθήκες

Πραγματοποιήσαμε επίσης πειράματα με βίντεο από μη εργαστηριακές συνθήκες και διεξήγαμε μια επιπλέον μελέτη χρηστών με 25 συμμετέχοντες. Αυτή τη φορά, αξιολογήσαμε μόνο τον οπτικό ρεαλισμό, λόγω των προαναφερθέντων περιορισμών. Χρησιμοποιήσαμε 2 ομιλητές από το σύνολο δεδομένων HDTF, ακολουθώντας τον ίδιο πρωτόκολλο όπως πριν. Παρουσιάζουμε τα αποτελέσματα της μελέτης χρηστών στον Πίνακα 5, και τις αντίστοιχες συγκρίσεις στο Σχήμα 6.4.

Οι χρήστες προτίμησαν σημαντικά το NEUTART έναντι του Wav2Lip και του SadTalker. Παρόλα αυτά το VideoReTalking προτιμήθηκε έναντι του NEUTART, αν και το μοντέλο μας δείχνει να δίνει υποσχόμενα αποτελέσματα. Ένα ποιοτικό συυπέρασμα που μπορεί να συναχθεί από τα παραπάνω πειράματα είναι ότι το NEUTART είναι σε θέση να δημιουργήσει εξαιρετικά ρεαλιστικά βίντεο, όταν εκπαιδεύεται σε δεδομένα κατάλληλης ποιότητας. Αντίθετα, οι συγκριθείσες few-shot μέθοδοι θυσιάζουν κάποια πτυχή της ποιότητας παραγωγής προκειμένου να είναι σε θέση να λειτουργήσουν σε μη περιορισμένες συνθήκες.

7 Συμπεράσματα

Σύνοψη

Η παρούσα Διπλωματική Εργασία παραθέτει τη σχετική βιβλιογραφία από τις περιοχές τη σύνθεσης φωνής, της μοντελοποίησης προσώπων, και της σύνθεσης ομιλούντων προσώπων, συνδυάζοντας επιμέρους μοντέλα προκειμένου να προτείνει μια καινοτόμα μέθοδο οπτικοακουστικής σύνθεσης ομιλίας. Η προτεινόμενη μέθοδος χρησιμοποιεί μετασχηματιστές για να μετατρέψει το κείμενο σε φασματογράφημα και 3Δ παραμέτρους προσώπου, χρησιμοποιώντας πολυτροπική εκπαίδευση. Το 3Δ βίντεο του ομιλούντος προσώπου έπειτα μετατρέπεται σε έγχρωμο και ενσωματώνεται σε κάποιο βίντεο αναφοράς, παράγοντας έτσι ένα φωτορεαλιστικό αποτέλεσμα. Η χρήση ακουστικών και οπτικών συναρτήσεων σφάλματος, και ιδιαίτερα οι συναρτήσεις όπως το σφάλμα ανάγνωσης χειλιών, χαρίζουν στο μοντέλο τη δυνατότητα σύνθεσης βίντεο με υψηλό ρεαλισμό, ιδίως ως προς την άρθρωση του στόματος κατά την ομιλία. Τα πειράματα και οι μελέτες που εκτελέστηκαν αποδεικνύουν την υπεροχή του μοντέλου μας όταν εκπαιδεύεται σε κατάλληλα σύνολα δεδομένων.

Μελλοντικές Επεκτάσεις

Το μοντέλο μας δεν καταφέρνει πάντα να παράξει ποιοτικά αποτελέσματα όταν εκπαιδεύεται σε δεδομένα που είτε δεν έχουν κατάλληλη ποιότητα ήχου, είτε έχουν αυτόματες απομαγνητοφωνήσεις, οι οποίες εισάγουν λάθη στη διαδικασία μάθησης για το σύστημα σύνθεσης φωνής. Συνεπώς, η αξιοποίηση κάποιας εύρωστης αρχιτεκτονικής κλωνοποίησης φωνής ενδεχομένως να βελτίωνε την ικανότητα προσαρμογής του σε θορυβώδη δεδομένα. Επιπλέον βελτιώσεις μπορούν να επιτευχθούν με τη βελτιστοποίηση της ταχύτητας της φωτορεαλιστικής μονάδας. Ακόμη, η εκπαίδευση από άκρο σε άκρο και των δύο μονάδων μπορεί να ενισχύσει περισσότερο το ρεαλισμό των συνθετικών βίντεο.

Ηθικά Ζητήματα

Θα θέλαμε να επισημάνουμε πως αν και τα συστήματα για φωτορεαλιστική σύνθεση ομιλίας μπορούν να έχουν πολύ θετικά αποτελέσματα σε διάφορες εφαρμογές, όπως η ψυχαγωγία, οι εικονικοί βοηθοί, ή τα εργαλεία προσβασιμότητας, υπάρχει ο κίνδυνος κακόβουλης χρήσης τους (Chesney and Citron 2019; Diakopoulos and Johnson 2021; Yadlin-Segal and Oppenheim 2021). Ένας κακόβουλος χρήστης μπορεί να συνθέσει βίντεο ατόμων χωρίς τη συναίνεσή τους, με σκοπό να παραπληροφορήσει ή να προσβάλει. Πιστεύουμε πως οι ερευνητές του κλάδου πρέπει να είναι ευαισθητοποιημένοι ως προς τα σχετικά ηθικά ζητήματα, καθώς και να συμβάλουν στην ανάπτυξη συστημάτων ανίχνευσης πλαστών πολυμέσων. Από μεριάς μας, δημοσιεύουμε τον πηγαίο κώδικα του συστήματος υπό μια ηθική άδεια, επιτρέποντας την ελεύθερη χρήση με την προϋπόθεση τήρησης της ηθικής δεοντολογίας.

Chapter 1

Introduction

Contents

Preface

In the early days of artificial intelligence, most models were focused on simple data analysis, performing either classification or regression. However, both the vast availability of data and the progress of accelerated hardware computing platforms has given rise to *deep learning* (LeCun et al. 2015). Deep learning consists of trainable models that have many processing layers, and has revolutionized data modeling. Nowadays, generative AI has demonstrated impressive capabilities in creating text, images, sound, and even entire 3D scenes. Countless software products that use such models are deployed to the cloud and have quickly gained thousands, or even millions of users (Zhang et al. 2023). Apart from its mainstream success in generating media, generative modeling is also useful in data augmentation, manipulation of high-dimensional distributions, model-based reinforcement learning, and semi-supervised learning.

Speech synthesis has been one of the pronounced successes of generative AI. In general, speech synthesis is the process of generating human-like speech, usually from a text input. It has captured the interest of researchers for decades, and nowadays it is even more popular, with the ubiquity of natural language conditional models. Text-to-Speech synthesis sits at the crossroads of linguistics, acoustics, and engineering, and has successfully adapted to the advent of deep learning and generative modeling. The first TTS attempts used simple concatenation of sounds in order to form the spoken text (Sagisaka et al. 1992). Obviously, the generated speech lacked naturalness and correct prosody. Since then, neural models have revolutionized speech generation. State-of-the-art TTS systems' output is almost indistinguishable from real human speech (Tan et al. 2021).

While the sound of speech is its most important aspect, the visual component is equally essential for conveying tone, emotion, and meaning. In this Diploma Thesis, we are targeting text-driven talking face generation, which is the process of generating the audio and video of a talking human character, with realistic and synchronized lip movements. Talking face generation aims at synthesizing videos of talking humans, with consistent and synced audio and visual streams. The field has many applications in entertainment, education, and virtual assistants.

The problem we are targeting is by definition multimodal and requires alignments between at least two modalities (text, audio, and video). Extending plain speech synthesis to audiovisual speech requires a deep understanding of the speech generation pipeline. Speech synthesis is inherently a one-to-many mapping, because a natural language phrase can be uttered in many ways, with different prosody, speed, intensity, and emotion. Of course, multi-speaker TTS is even more challenging, and needs to incorporate information about each speaker's distinct speaking style. The visual aspect of talking face generation is also one-to-many, as each phrase can be uttered with various facial expressions. Deep generative models can address the one-to-many problem by modeling the distribution of outputs that correspond to a given input. Thus, we present a short introduction of deep learning and generative modeling before addressing the main topic of this Thesis.

1.1 Deep learning

Machine learning models can be classified into two major categories based on their approach of describing the learned data. On the one hand, there are *discriminative* models, which describe the distribution $p_\theta(y|x)$, that is to say the distribution that predicts a label or value y based on some dataset instance x, and using parameters θ. On the other hand, *generative* models can either describe $p_\theta(x, y)$ or $p_\theta(x)$, and can be sampled to generate new data that could be instances of the training dataset (Goodfellow et al. 2016). In this Section we present the most prominent generative models and some of their applications.

To illustrate more clearly the discrepancy between the two approaches, let us consider a dataset of speech recordings x, along with their text transcriptions y. We can use this dataset to train models for two different tasks:

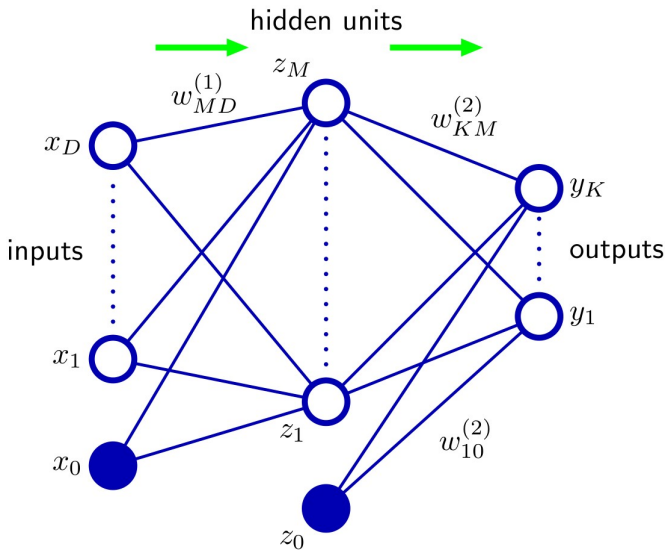


Figure 1.1: Diagram for a neural network of two layers. The input, hidden (intermediate), and output variables are represented by nodes, while the weight parameters are represented by links between the nodes. Also, the bias parameters are denoted by links coming from additional input and hidden variables x_0 and z_0. Arrows denote the direction of information flow through the network during computation of the outputs (inference). Figure from Bishop and Nasrabadi (2006).

- A discriminative model for Automatic Speech Recognition, whose goal is to predict: $p_\theta(\texttt{text}|\texttt{speech})$
- A generative model for Text-to-Speech, thus modeling the conditional distribution: $p_\theta(\texttt{speech}|\texttt{text})$

1.1.1 Feedforward Neural Networks

The feedforward neural network is the simplest neural architecture, consisting of a series of functional transformations. Each transformation depends on learnable parameters and can be modeled as a neuron in a graphical model such as the one presented in Figure 1.1. Suppose a network input $\vec{x} = [x_1, ..., x_D]^T$. Each input neuron j multiplies its input x_i with a weight w_{ji}, adds a bias w_{j0} and performs a (usually nonlinear) activation function h to the result (Bishop and Nasrabadi 2006). Formally, each neuron's transformation can be written as:

$$z_j = h\left(\sum_i w_{ji}x_i + w_{j0}\right) \tag{1.1}$$

By using multiple neurons in a layer and chaining more layers of different depths, we can build parametric models that can be trained to model some output quantity based on the network input. Increasing the number of parameters improves the network's ability to model complex distributions. Such networks with many layers are called *deep*, thus giving the name to the field of deep learning.

1.1.2 Neural Network Training

A network's learnable parameters are iteratively optimized by gradually minimizing some error or *loss* function. During training, the inputs are processed by the network, producing the outputs. Then, the outputs are compared with the desired outputs in order to calculate the error gradients, which are used to optimize the trainable parameters. The gradient descent algorithm can be

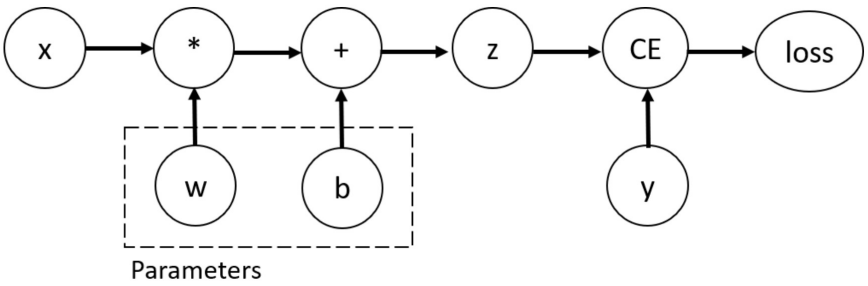


Figure 1.2: An computational graph example for the operation $\mathcal{L}(y, wx + b)$, where $\mathcal{L}$ is the cross-entropy loss function (`CE`). The network parameters that need to optimized are the weight w and bias b. The PyTorch framework stores the gradients required for the loss optimization, for instance $\frac{\partial z}{\partial w} = x$ and $\frac{\partial \mathcal{L}}{\partial z}$, which is a predefined closed form expression, derived by directly differentiating the cross-entropy loss.

used to update the network parameters by following the path in the parameter space which corresponds to the maximum local steepness. A commonly used version of the algorithm is *Stochastic Gradient Descent* (SGD), which is superior in terms of computational efficiency, since it operates on mini-batches of data instead of the entire dataset. SGD iteratively adjusts model parameters in the direction of the gradient estimated from each mini-batch, allowing for faster convergence and often escaping from local minima. The weight update expression can be written, for a time step t and a learned parameter w:

$$w^{t+1} = w^t - \eta \frac{\partial \mathcal{L}(w)}{\partial w} \tag{1.2}$$

where $\mathcal{L}$ is a differentiable loss function controlled by a set of parameters, and η is the learning rate. This equation can be easily generalized and written in vector or matrix notation for more complex network parameters.

The backpropagation algorithm (LeCun et al. 2015) calculates the aforementioned error gradients by efficiently applying the differentiation chain rule. Thus, backpropagation and gradient descent work together in order to optimize a network's parameters. The former facilitates efficient computation of gradients, enabling the latter to iteratively update model parameters, driving the network towards convergence.

Deep learning software like PyTorch (Paszke et al. 2019), which was extensively used in this Diploma Thesis, can perform optimization using gradient descent thanks to their implementation of automatic differentiation in computational graphs. We provide an illustration of a computational graph in Figure 1.2.

1.1.3 Sequential Architectures

Apart from neural networks that operate on instances of data without a temporal dimension, like vectors or images, data such as speech or natural language can be processed by neural networks that need to model their *sequential* nature. Formally, the problem of processing sequential data can be formulated as finding a function $\mathcal{F}$ to map a sequence of inputs $x_{1:N}$ to a sequence of outputs $y_{1:M}$:

$$y_{1:M} = \mathcal{F}(x_{1:N}) \tag{1.3}$$

Notice that the input and output sequences do not need to have the same length. For instance, a machine translation task needs to map sentences from one language to another, most likely using a different numbers of words in each one.

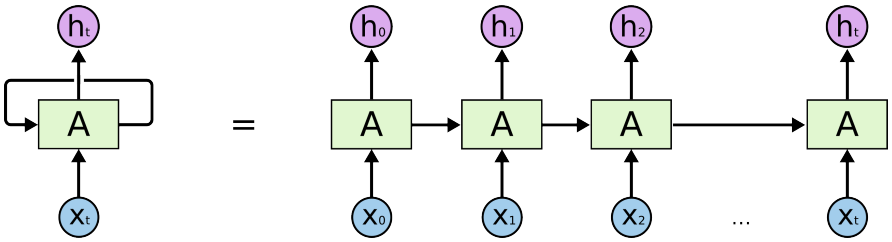


Figure 1.3: A neural network with a feedback loop (left), unrolled in time for visualization of the parameter sharing mechanism (right). Once unrolled, we can clearly see that each input x_t is processed by the same network parameters A. This particular network processes the sequential inputs and outputs a value h_t in each time step, while also keeping information about previous inputs.

Recurrent Neural Networks

One approach to sequential modeling would be to add a feedback loop into a neural network, thus converting it to a dynamical system (LeCun et al. 2015). The recurrence is essential for the ability to model sequential data, since the feedback loop would provide information about past inputs at each time step. While such *Recurrent Neural Networks* (RNNs) have been successful in applications such as speech and language processing (Yao et al. 2013), their main drawback is evident once we look at the example RNN of Figure 1.3.

Namely, the feedback loop in the computation graph acts as a bottleneck for information. Thus, recurrent models suffer from the *long-term context* problem, since they are not able to effectively retain all the relevant information that they have been presented with.

Long-Short Term Memory (LSTM) networks (Hochreiter and Schmidhuber 1997) attempt to mitigate the long-term context problem of RNNs, by carefully gating the information flow across time steps. For more details on recurrent models, we refer to the detailed article by Sherstinsky (2020).

The Transformer

Vaswani et al. (2017) proposed the very successful transformer architecture for sequence modeling, which tackles the long-term context problem by eliminating recurrence. In contrast, transformers are autoencoders that process sequential inputs by modeling relationships between elements via scaled dot-product attention or *self-attention*. Residual connections are also utilized (He et al. 2016), meaning that the input of each network's layer is added back to the output.

The mechanism of self-attention aims at attending on the inputs according to their importance. For that, the vector inputs are projected to learned low-dimensional query, key, and value subspaces, with the query and key subspaces being of dimensionality p. Formally, each input $\vec{x}_i \in \mathbb{R}^d$ is projected to the vectors $\vec{q}_i = \mathbf{W}_Q \vec{x}_i$, $\vec{k}_i = \mathbf{W}_K \vec{x}_i$, and $\vec{v}_i = \mathbf{W}_V \vec{x}_i$, with $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{p \times d}$ and $\mathbf{W}_V \in \mathbb{R}^{r \times d}$. The three $\mathbf{W}$ matrices are trainable parameters that learn the projection to the aforementioned spaces. The attention score $\vec{y}_i \in \mathbb{R}^r$ is then calculated as:

$$\vec{y}_i = \sum_j \text{softmax}\left(\frac{1}{\sqrt{p}} \vec{q}_i^{\,T} \vec{k}_j\right) \vec{v}_j \tag{1.4}$$

Intuitively, the self-attention's output at each position is a weighted sum of all input transformations (values), interpolated using a similarity measure between the current input (query) and the other items in the sequence (keys). By combining all these vectors into matrices, for instance $\mathbf{X} = [\vec{x}_1^T, ..., \vec{x}_N^T]^T \in \mathbb{R}^{N \times d}$, the attention output matrix $\mathbf{Y} \in \mathbb{R}^{N \times r}$ can be neatly calculated using the matrices $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{N \times d}$, and $\mathbf{V} \in \mathbb{R}^{N \times r}$ of concatenated vectors:

$$\mathbf{Y} = \text{softmax}\left(\frac{1}{\sqrt{p}} \mathbf{Q}\mathbf{K}^T\right) \mathbf{V} \tag{1.5}$$

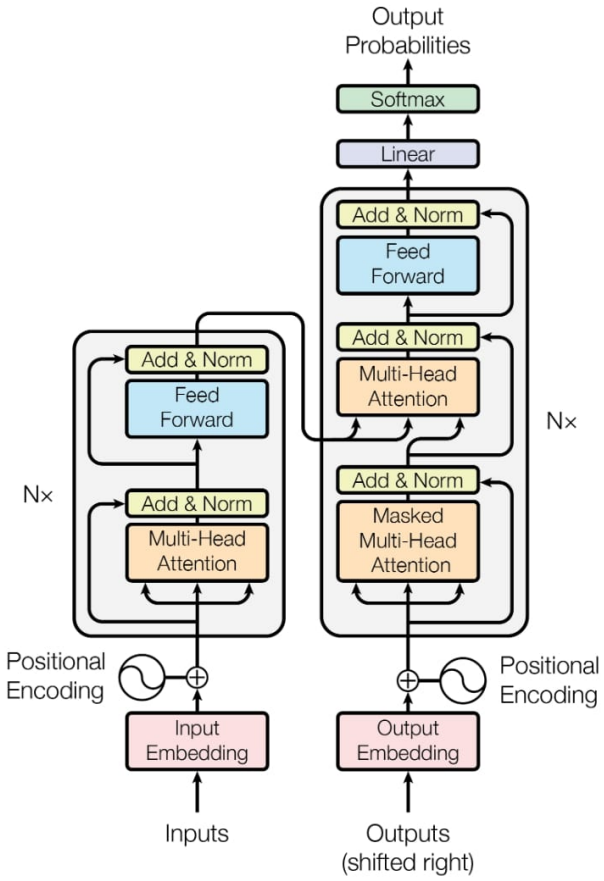


Figure 1.4: The transformer encoder-decoder architecture. An encoder stack processes the input sequence using multi-head attention and feedforward layers, creating an intermediate representation, which is fed into each decode module. Notice the residual connections at the output of each sublayer, before normalization. The non-recurrent nature of transformers requires the explicit integration of temporal information, which is implemented by adding positional encodings to each sequence item. Figure from Vaswani et al. (2017).

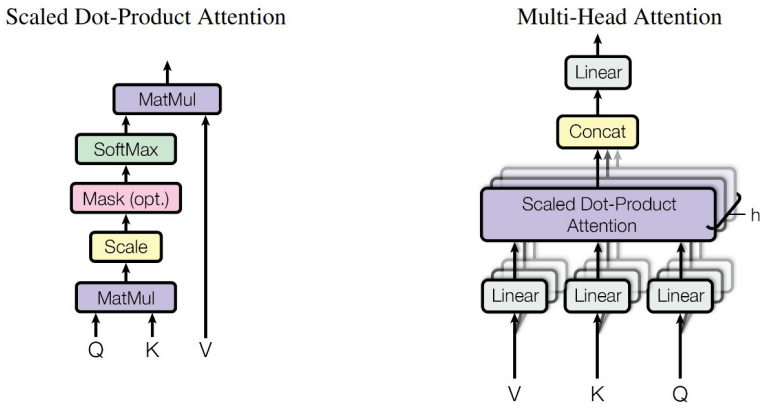


Figure 1.5: Visualization of the scaled dot-product attention computational pipeline (left), directly implementing Equation (1.5), as well as the multi-head attention (right). Multi-head attention is performed by first projecting the queries, keys, and values into different learned subspaces for each head. Figure from Vaswani et al. (2017).

Instead of performing a single attention computation, the authors propose multi-head attention, which performs the computation of Equation (1.5) h times, in order to model different types of dependencies. The block diagram of a transformer model, as well as the multi-head attention mechanism, are presented in Figures 1.4 and 1.5.

1.2 Generative Modeling

1.2.1 Generative Adversarial Networks

One major breakthrough in deep generative modeling came with *Generative Adversarial Networks* (GANs) (Goodfellow et al. 2014), which are devised from the game theory scenario of a minimax game. Such a game is played by two players and in all states the reward of one player is the negative of reward of the other. In GANs, the players are two competing neural network models that are trained jointly: a generative network $\mathcal{G}$ that models the data distribution $p(x)$, and a discriminative network $\mathcal{D}$ that estimates the probability that a sample came from the training data rather than $\mathcal{G}$. The training procedure aims at $\mathcal{G}$ learning the data distribution well enough to "deceive" $\mathcal{D}$.

More formally, the discriminator and the generator play a two-player minimax game with value function $V(\mathcal{G}, \mathcal{D})$. A prior $p(z)$ is defined on noise variables z, which act as the the generator inputs. The probability that a sample x comes from the data distribution rather than the generator is $\mathcal{D}(x)$. Thus, the value function can be written as:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{x \sim p(x)}[\log \mathcal{D}(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - \mathcal{D}(\mathcal{G}(z)))]$$

Adversarial training is done in two steps. First, random samples x from the training set are passed through the discriminator, whose goal is for $\mathcal{D}(x)$ to be near 1, meaning that it recognizes the samples as real. Then, the generator generates a fake sample using an input z randomly sampled from $p(z)$. The discriminator aims to make $\mathcal{D}(\mathcal{G}(z))$ approach 0, while the generator strives to make the same quantity approach 1. See Figure 1.6 for an overview of the training process.

However, Goodfellow (2016) notices that early in learning, when $\mathcal{G}$ is untrained, $\mathcal{D}$ can easily reject generated samples because they are clearly different from the training data, thus not providing strong gradients to optimize $\mathcal{G}$. In order to have stronger gradients early in learning, $\mathcal{G}$ can be trained with the equivalent objective of maximizing $\log \mathcal{D}(\mathcal{G}(z))$, which leads to the same fixed point of the game dynamics. Since the discriminator operates under the assumption that half of the inputs are real and half are fake, the Nash equilibrium of the game corresponds the point where $\mathcal{G}(z) = p(x)$ and $\mathcal{D}(x) = 0.5$, for any x. This means that the generator has perfectly captured the training distribution, and the discriminator randomly decides if a sample is real or fake with equal probability.

Of course, unconditional models are useful for general sampling from a distribution. In order to obtain specific examples, we need to add some kind of conditioning (Mirza and Osindero 2014). All of the above definitions can be easily augmented for the conditional case, where the conditioning can be anything from a text representation to an image.

The original authors claim that since the generator never actually processes the training data, GANs are not prone to overfitting. Nevertheless, they exhibit a number of drawbacks, including mode collapse, where the generator maps several different noise inputs to the same output, severely limiting the sampling diversity. Instability in training is also very common, where the optimizer can either achieve a poor local minimum, reach no minimum at all, or completely diverge (Goodfellow 2016; Oussidi and Elhassouny 2018).

Applications

While difficult to train, GANs have been successfully used for generative modeling since their invention. In the case of image generation, they produce much sharper images than previous

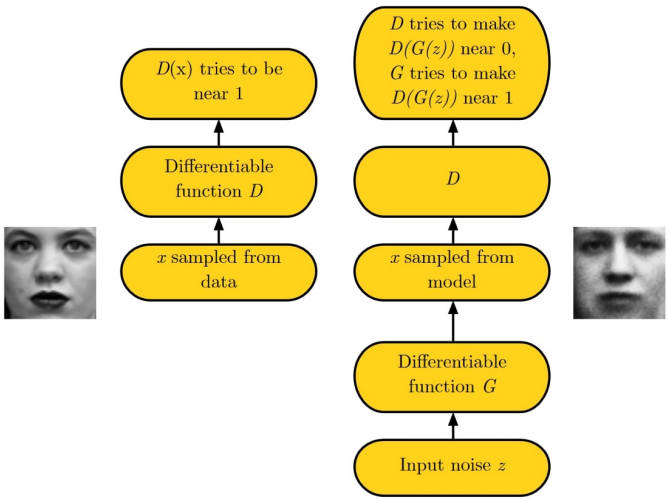


Figure 1.6: Visualization of GAN training process on a dataset of human faces. In this Figure, the generator and discriminator are symbolized as G and D, respectively, while the dataset images are symbolized as x. The GAN framework pits two adversaries against each other in a game which plays out in two scenarios. In one scenario, example face images x are randomly sampled from the training set and used as input for the discriminator D. The goal of D is to output the probability that its input is real, which translates to $D(x)$ being near 1. In the second scenario, inputs z to the generator are randomly sampled. Then, the discriminator then receives input $G(z)$, a fake face image created by the generator. In this scenario, both players participate. D backpropagates a loss computed with $D(G(z)) = 0$ as a target value, while G backpropagates a loss computed with $D(G(z)) = 1$ as ground truth. Figure from Goodfellow (2016).

approaches. One seminal adaptation of the GAN architecture for image synthesis is Deep Convolutional GAN (DCGAN), by Radford et al. (2015), which uses solely convolutions to convert a random noise vector into an image. The information flow during the generator's inference is presented in Figure 1.7.

Other notable works are StyleGAN and StyleGAN2 (Karras et al. 2019; Karras et al. 2020), that use the adversarial framework for arbitrary style transfer in images. Their proposed method differs from the original GAN and DCGAN, due to the architectural innovations that they introduced. These innovations allow the model to separate the high-level features as well as offer stochastic variation. The generation does not start from a random noise sample, but rather from a learned constant input, which is processed by convolutional and normalization layers to generate an image. The stochastic variation of detail is generated by introducing uncorrelated Gaussian noise inputs that are infused into the feature maps.

More recently, the unprecedented success of text-conditioned image synthesis has highlighted one of the neglected advantages of GANs, which is their computational efficiency. While diffusion models (Sohl-Dickstein et al. 2015; Ho et al. 2020) have dominated the area of image synthesis due to their ability to generate incredibly realistic samples, the diffusion process requires iterative forward passes through neural networks, often in the order of hundreds of iterations. On the other hand, an adversarially trained network is able to generate samples in only one forward pass. Kang et al. (2023) leverage a pretrained vision-language model (Radford et al. 2021) and follow a progressive upsampling approach to produce high-resolution images, generated in orders of magnitude less time than the state-of-the-art diffusion image generators.

In the context of human faces, Generative Interpretable Faces (Ghosh et al. 2020) is a StyleGAN2-based framework that generates photorealistic face images with explicit control over face geometry and style parameters. It uses 3D parameters as the face geometry and expression conditioning, and a style vector for other factors such as hairstyle and background. They authors

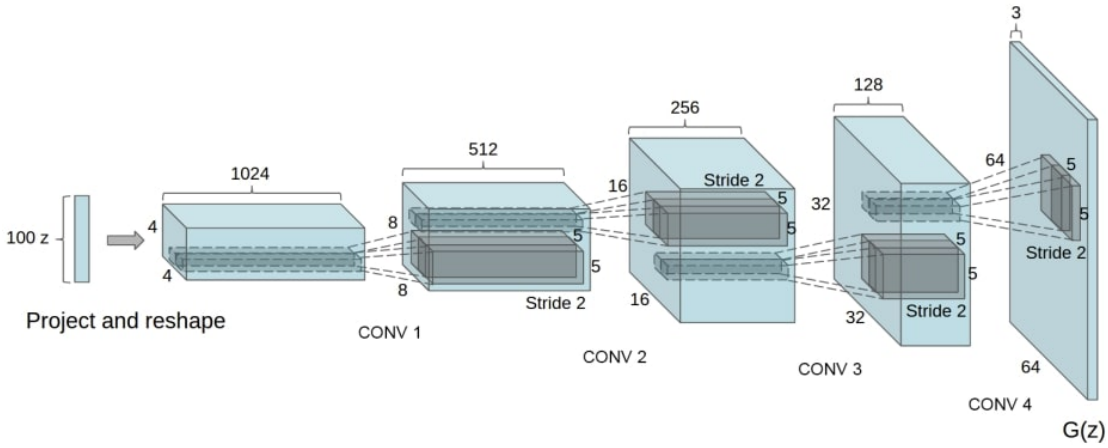


Figure 1.7: The generator's inference process in the Deep Convolutional GAN architecture. A noise vector $\vec{z} \in \mathbb{R}^{100}$ is sampled from a uniform distribution and projected to a small spatial extent convolutional representation with many feature maps. Then, a series of four strided convolutions convert this high level representation into a colored pixel image. Figure from Radford et al. (2015).

also highlight the issue of condition cross-talk, meaning that conditions that are presumed to be independent tend to influence each other.

1.2.2 Diffusion Models

Sohl-Dickstein et al. (2015) and Ho et al. (2020) proposed a new way of learning distributions, inspired by non-equilibrium statistical physics. They devised the idea that a data distribution can be learned by slowly destroying its structure through a diffusion process (called *forward* process), and learning how to reverse it. Their experiments produced models that were successful at generating data, most notably for image generation and image inpainting.

Suppose a generative model aiming to learn the distribution of $\vec{x} \in \mathbb{R}^d$. A raw sample from that distribution would be $\vec{x}_0$, which is corrupted with noise and apply noise corruption in T steps. The sample after the t-th noise addition is denoted as $\vec{x}_t$. The noise addition is governed by the forward Gaussian process q with schedule β_t, satisfying the Markovian structure:

$$q(\vec{x}_t|\vec{x}_{t-1}) = \mathcal{N}(\vec{x}_t; \sqrt{1-\beta_t}\vec{x}_{t-1}, \beta_t \mathbf{I})$$

The denoising distribution is $q(\vec{x}_{t-1}|\vec{x}_t)$, which is intractable. However, provided that β_t is small, it can be approximated by a Gaussian distribution. We can now define the *reverse* process p, with $p(\vec{x}_T) = \mathcal{N}(\vec{x}_T; \vec{0}, \mathbf{I})$, since after T noising steps the original data is lost, leaving only random noise.

$$p_\theta(\vec{x}_{t-1}|\vec{x}_t) = \mathcal{N}(\vec{x}_{t-1}; \mu_\theta(\vec{x}_t, t), \sigma_t^2 \mathbf{I}) \implies p_\theta(\vec{x}_{0:T}) = p(\vec{x}_T) \prod_{t=1}^{T} p_\theta(\vec{x}_{t-1}|\vec{x}_t)$$

The function for the mean, $\mu_\theta(\vec{x}_t, t)$, is computed by a trainable network, while the standard deviation can either be trainable or constant. In the original formulation, σ_t^2 is set equal to β_t. See Figure 1.8 for a visualization of the process.

Applications

Rombach et al. (2022) leveraged the powerful encoder-decoder architecture in order to perform the diffusion process in a compact and computationally efficient latent space. Popular text-to-image generative models such as DALL·E and Stable Diffusion are implemented in a latent space.

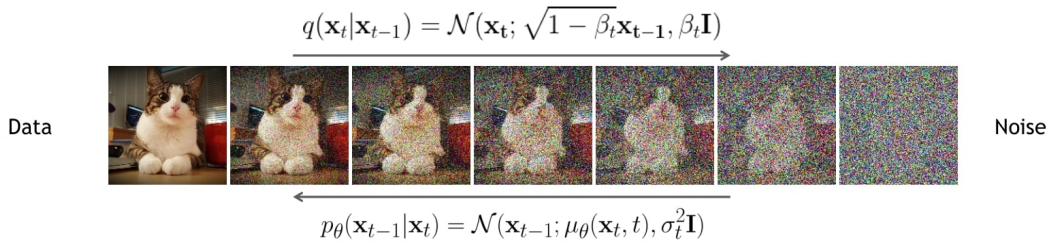


Figure 1.8: Visualization of diffusion on an image sample. The forward diffusion is a Gaussian process that slowly corrupts the image with noise, until nothing but a random sample from a normal distribution is left. The reverse process performs iterative denoising by estimating the mean value of the noise at that particular time step and subtracting it. Figure from Kreis et al. (2022).

Learning consists of two stages: a perceptual compression stage maps the image into the latent space, then a generative model learns the semantic and conceptual composition of the data. The model is realized as a time-conditional U-Net (Ronneberger et al. 2015). For conditional generation, an intermediate representation of the condition is mapped into the U-Net via a cross-attention layer.

We note once again that sampling from a diffusion model requires T denoising operations, which is obviously computationally expensive, as T is usually in the order of 10^3. Denoising Diffusion GANs (Xiao et al. 2021) attempt to combine diffusion models' high quality samples with GANs' sampling speed, constructing a network that leverages the generative abilities of GANs to denoise images in only 4 steps. Many other approaches have been attempted to accelerate inference on diffusion models. For instance, Denoising Diffusion Implicit Models (Song et al. 2020) modify the Markovian structure of diffusion models, by modeling the forward process as $q(\vec{x}_t|\vec{x}_{t-1}, \vec{x}_0)$. This leads to the same training objective, but allows for accelerated sampling. We refer to Yang et al. (2022) for a complete survey of diffusion models' theory and applications.

Kim et al. (2022) successfully use diffusion to implement face swapping framework that uses facial guidance from external models, to synthesize images with a given source identity, while preserving the target image's attributes. Video diffusion models, by Ho et al. (2022) extend the 2D convolutions into 3D space-only convolutions for videos. In order to decouple the space and time dimensions, the time axis is treated as batch axis in spatial convolutions, and spatial axes are treated as batch the temporal attention blocks. The authors claim that this disentanglement is useful for also applying the model to images.

One aspect of generative modeling that is related to audiovisual speech synthesis is the generation of human motion sequences, mostly used for entertainment, gaming or robotics control. The produced motions must be both lifelike and temporally coherent. Zhang et al. (2022a) and Tevet et al. (2022) proposed the first text-driven motion generation frameworks that use diffusion. The former offers body-part time-dependent control over the sequence, while the latter incorporates geometric losses and is more lightweight. Both approaches model the human motion as a sequence of geometric pose states, and use the transformer architecture (Vaswani et al. 2017) instead of the U-Net that is almost ubiquitous in image generation. The conditioning mechanism is similar to image generation, since the text describes the whole motion sequence, so the frameworks use CLIP (Radford et al. 2021) in order to extract text features for the generation guidance. Chen et al. (2022) use diffusion on a latent space to perform the same task.

Nair et al. (2022) explore image generation with multi-modal conditioning. They experiment with combined conditions such as text, image masks, and image sketches. Assuming independent modalities, they derive an expression for performing exact sampling through a score-based approach. In the context of audiovisual works, MM-Diffusion, by Ruan et al. (2022) generates aligned audio-video pairs. They use independent forward processes for each modality, with a shared schedule, and a coupled U-Net for denoising. Since they experiment mostly with land-

scape and dancing datasets, they argue that data is temporally redundant, thus they propose Random-Shift based multi-modal attention for efficiency.

1.3 Contributions

As already mentioned, the main topic of this Diploma Thesis is talking face generation. Our work aims to leverage the advancements in deep learning that were outlined in this introductory Chapter, and use them in the context of photorealistic audiovisual speech synthesis. Towards that end, we build upon both speech synthesis and face modeling techniques, which is thoroughly explored in the following Chapters. Our contributions to the field of photorealistic talking head generation can be briefly summarized as follows:

- We introduce the first, to the best of our knowledge, text-driven, photorealistic audiovisual speech synthesizer that is genuinely bimodal and avoids the cascaded 2-stage approaches for audio and video synthesis adopted by previous methods.
- We propose a novel joint modeling of acoustic and 3D visual elements in a learned feature space, which captures the complex interplay between audio and visual streams. Our experiments show that this can increase the perceived realism and plausibility of the final synthetic result.
- We adopt an accurate 3D representation for the synthesis of visual speech and combine it with state-of-the-art photorealistic video synthesis based on conditional generative adversarial networks. This allows us to blend the synthesized facial motions that match the input text with various scenes in a photorealistic manner, paving the way for a multitude of extended capabilities for AI-based video synthesis.
- We conduct qualitative and quantitative experiments, as well as user and ablation studies to evaluate our method and compare it with recent state-of-the-art methods. The experiments demonstrate the effectiveness and advantages of our method, which surpasses previous methods using lab-recorded datasets, and also achieves particularly promising results in challenging in-the-wild scenes.
- We make the source code of our method publicly available at the project's website, under an ethical license: Milis (2023).

1.4 Organization

Having already introduced the reader to the concept of audiovisual speech, the field of deep learning, as well as some aspects of generative modeling, the rest of this Diploma Thesis is organized as follows:

- In Chapter 2 we analyze the mechanism and related work on of speech synthesis.
- In Chapter 3 we present the background and related work in human face modeling.
- In Chapter 4, we build on the previous two Chapters and explore the related bibliography on talking face generation.
- Chapter 5 thoroughly analyzes our proposed text-driven audiovisual model for talking face generation.
- Chapter 6 presents the experiments we conducted with the proposed model.
- Finally, Chapter 7 offers a short overview of our results, discusses potential future work, and raises some relevant ethical considerations.

1.5 Notation

In this Section, we briefly outline the notation to which we adhere when presenting mathematical expressions, unless specified otherwise.

- We represent scalars with lowercase letters, for example a certain frequency $f \in \mathbb{R}$.
- Vectors are represented with arrows, such as a feature vector $\vec{x} \in \mathbb{R}^d$.
- We denote images, matrices, or tensors with bold capital letters, such as an RGB image $\mathbf{I} \in \mathbb{R}^{W \times H \times 3}$.
- Sequences and their items are written with subscripts indicating their indexing, for instance a sequence of N feature vectors would be $\vec{x}_{1:N}$, with its n-th element being $\vec{x}_n \in \mathbb{R}^d$.
- Finally, neural networks and loss functions are written with calligraphic capital letters, for example a trainable network $\mathcal{F}$.

Chapter 2

Speech Synthesis

Contents

Preface

In this Chapter we explore the area of *Text-to-Speech* (TTS) technology, which maps plain text to realistic and dynamic voices. TTS sits at the crossroads of linguistics, machine learning, and digital synthesis. Its applications are numerous, from enhancing accessibility for visually impaired individuals, to virtual assistants and audiobook narration. We first cover some of the theoretical background behind speech synthesis technology, focusing on the preliminaries of the system that we used in this Diploma Thesis. Then, we present some relevant works, showcasing the evolution of the field from simple methods to neural networks.

2.1 Preliminaries

In this Section we cover the preliminaries of TTS synthesis. Let us consider a TTS system. Its input would be raw text as a sequence of characters, while the desired output would be the waveform of an audio signal that resembles a person uttering the input text. What differentiates TTS models from other systems is that their input is not a signal in itself, it has to be processed and then converted to a form suitable for computation.

2.1.1 Phonetic Modeling

As Rabiner and Schafer (2010) describe, the text needs to be normalized, meaning that capitalization, symbols and punctuation are either removed or appropriately replaced. For instance, numbers or symbols are spelled out. Then, the text has to be mapped to a sequence of *phonemes*, which are the smallest distinct units of language that can be used to compose words. This phonetic mapping can be done via a phonetic dictionary for known words, or use some predefined or learned mapping for unknown pronunciations. In this thesis, the Carnegie Mellon Pronouncing Dictionary (CMUdict) was used, which is both widely used and open-source.

The CMUdict is a pronunciation dictionary for North American English that contains over 134,000 words and their pronunciations in the ARPAbet phoneme set. Stress in vowels is denoted by the lexical stress markers `0-2`:

- `0`: No stress
- `1`: Primary stress
- `2`: Secondary stress

The phoneme set has 39 phonemes, not counting the variations due to lexical stress. For instance, the phrase "a pronunciation example" would be transcribed to:

```
AH0  P R OW0 N AH2 N S IY0 EY1 SH AH0 N  IH0 G Z AE1 M P AH0 L
```

The entire ARPAbet phoneme set with example transcriptions is presented in Table 2.1.

However, we should note that the phonetic split is not always trivial. While some languages' pronunciation can be easily inferred from their spelling, this is not the case for English. The phoneme mapping step should take into account the cases of homographs (different words that are spelled out the same), as well as *out-of-vocabulary* (OOV) words. In the case of homographs, they can be disambiguated using their part-of-speech in most cases. In contrast, OOVs require a phoneme sequence prediction, which can be done with a learned mapping.

2.1.2 Alignment

The training of TTS models requires paired audio waveforms with text transcriptions. Each audio sequence has to be aligned with the text's phoneme sequence, meaning that the model has to know exactly which audio segment contains each particular item of the phoneme sequence. See Figure 2.1 for a visualization of such a sequence alignment. Many TTS frameworks use the *spectrogram* as an intermediate representation, from which a *vocoder* can then synthesize the

Phoneme	Example	Transcription
AA	odd	AA D
AE	at	AE T
AH	hut	HH AH T
AO	ought	AO T
AW	cow	K AW
AY	hide	HH AY D
B	be	B IY
CH	cheese	CH IY Z
D	dee	D IY
DH	thee	DH IY
EH	Ed	EH D
ER	hurt	HH ER T
EY	ate	EY T
F	fee	F IY
G	green	G R IY N
HH	he	HH IY
IH	it	IH T
IY	eat	IY T
JH	gee	JH IY
K	key	K IY
L	lee	L IY
M	me	M IY
N	knee	N IY
NG	ping	P IH NG
OW	oat	OW T
OY	toy	T OY
P	pee	P IY
R	read	R IY D
S	sea	S IY
SH	she	SH IY
T	tea	T IY
TH	theta	TH EY T AH
UH	hood	HH UH D
UW	two	T UW
V	vee	V IY
W	we	W IY
Y	yield	Y IY L D
Z	zee	Z IY
ZH	seizure	S IY ZH ER

Table 2.1: Examples for each of the 39 phonemes of ARPAbet, spelled out in the CMU pronunciation dictionary. Vocal stress is ignored. Table from the Carnegie Mellon Speech Group website: http://www.speech.cs.cmu.edu/.

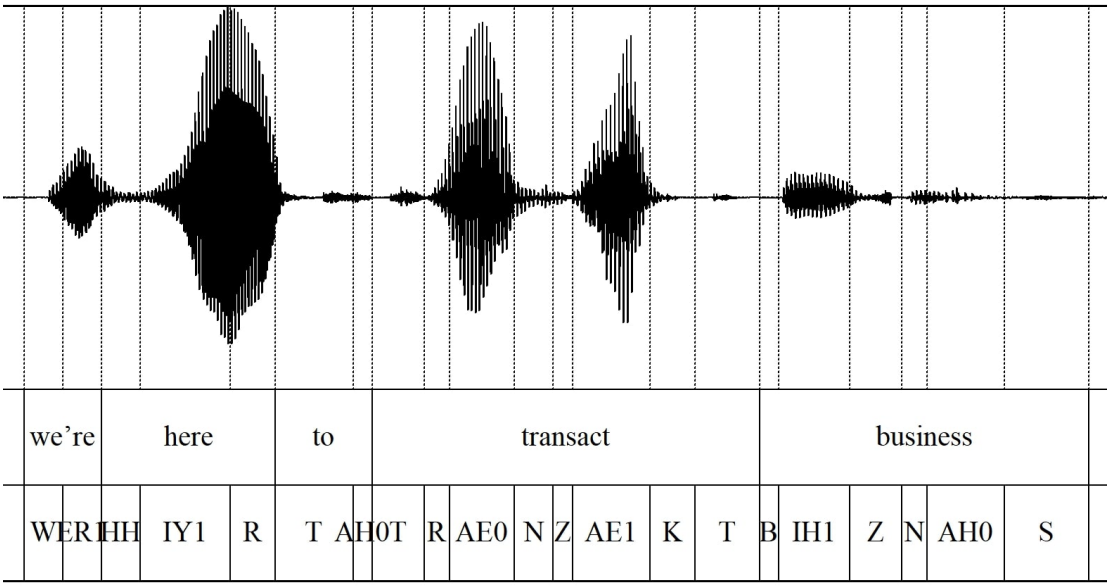


Figure 2.1: A visualization of alignment between audio, text, and phonemes from a sample of the TCD-TIMIT dataset (Harte and Gillen 2015). The alignment was performed using the Montreal Forced Aligner (McAuliffe et al. 2017) and the visualization is made in Praat (Boersma and Weenink 2009). Notice the high amplitude in vowels, especially those with primary stress.

speech waveform. The spectrogram contains information about the signal's frequency components and is usually represented in mel scale. Alternatively, an end-to-end system can be trained to directly map phoneme sequences to waveforms, without utilizing a vocoder. Deep learning-based TTS models initially tackled generation in an autoregressive manner, implying that they suffered from slow inference speed. In addition, they weren't robust enough and used to skip or repeat words. However, parallel approaches tackle both the speed and robustness problems.

2.1.3 The Spectrogram

The spectrogram contains information about the frequency content of a signal as it varies over time. Thus, it can be calculated by applying the Fourier transform to a sliding window over the signal, then concatenating the extracted frequency representations over time as a heatmap image. Formally, the spectrogram $\mathbf{F}$ of a digitally sampled signal $x[n]$, such as a speech waveform, is the magnitude of its *Short-Time Fourier Transform* (STFT).

$$\mathbf{F}(m,\omega) = |\mathrm{STFT}(m,\omega)|^2 \tag{2.1}$$

The STFT can be calculated using a window $w[n]$ that isolates the signal's content to a particular time frame.

$$\mathrm{STFT}\{x[n]\}(m,\omega) = \sum_n x[n]w[n-m]e^{-j\omega n} \tag{2.2}$$

An example of a speech signal's spectrogram is presented in Figure 2.2.

Instead of expressing the spectrogram in physical frequency scale, we can transform the representation from Hertz to the mel scale, which is more suitable for human perception. The mel scale m is a perceptual scale of pitches judged by listeners to be equal in distance from one another, extracted via experiments (Stevens et al. 1937), and can be expressed as a logarithmic transform of the standard frequency scale f.

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.3}$$

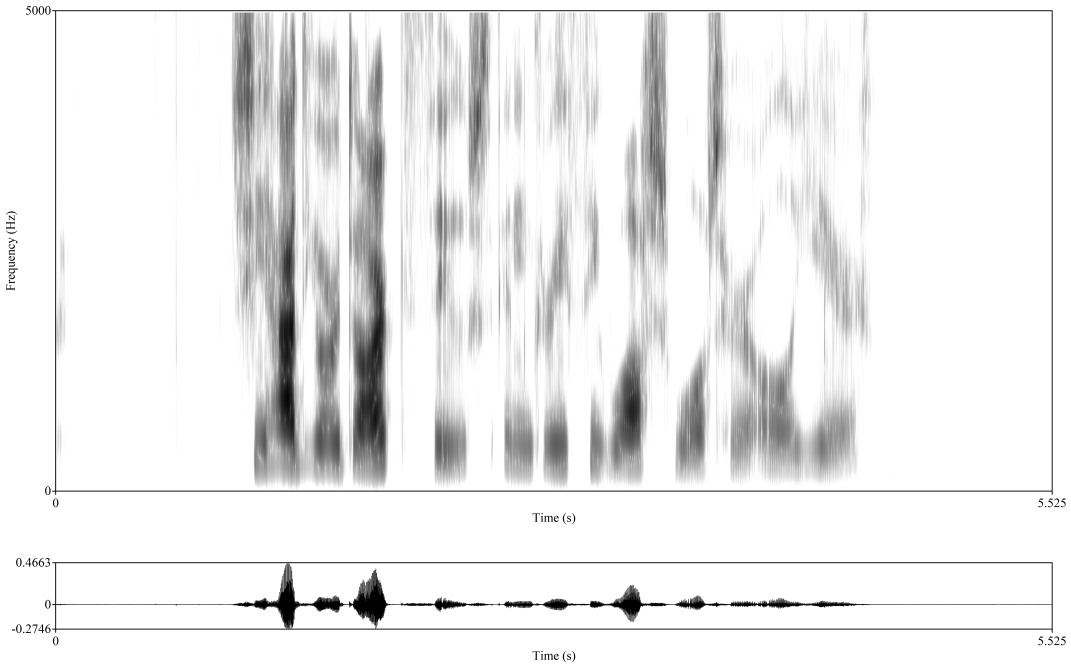


Figure 2.2: The spectrogram of the utterance "She had your dark suit in greasy wash water all year", spoken by a female subject. The speech waveform is also shown at the bottom. The sentence's vowels correspond to high amplitude in the waveform and high energy (darker) regions in lower frequencies. Again, the audio is from a sample of the TCD-TIMIT dataset, and the image is made in Praat.

While the STFT is an invertible transformation, the phase information is lost in the spectrogram. Thus, applying inverse Fourier transform in temporal windows of a spectrogram cannot reliably retrieve the original signal. This led to the development of neural models that perform spectrogram to waveform mapping. In the context of speech, this type of networks are referred to as vocoders.

The increasing presence of intelligent voice agents in everyday life has led to extensive research in the area of speech synthesis. In the next Section we explore the history as well as the most prominent recent TTS systems, which are useful for the development of our audiovisual talking face generation system.

2.2 Models

2.2.1 Older Approaches

The ideal approach to speech synthesis would be to simulate the human articulation mechanism such as the geometry and movement of the lips, tongue, glottis, and vocal tract. However, it is very difficult to model these articulation behaviors in practice, mainly due to difficulties in data collection. Therefore, approaches in articulatory synthesis such as the ones by Coker (1976) or Shadle and Damper (2002), did not prevail due to low quality generation.

Concatenative synthesis relies on the concatenation of speech units stored in a database. Those units may range from whole sentence to syllables that are recorded by voice actors. Moulines and Charpentier (1989) used diphones for effective concatenative synthesis. During inference, the concatenative TTS system searches speech units to match the given input text, and produces the speech waveform by concatenating these units together. Generally speaking, concatenative TTS can generate audio with high intelligibility and authentic timbre close to the original voice

actor (Sagisaka et al. 1992; Dutoit 1997). However, concatenative TTS requires huge recording database in order to cover all possible combinations of speech units for spoken words. Another drawback is that the generated voice is less natural and emotional, since concatenation can result in less smoothness in stress, emotion, or prosody. Nevertheless, such methods remained prominent until recently (Chalamandaris et al. 2010).

Later, TTS models employed a statistical parametric approach (Tokuda et al. 2000; Yoshimura 2002) to alleviate the drawbacks of unit concatenation, by first generating the necessary acoustic parameters, then recovering the speech waveform using a vocoder. It usually consists of a text analysis module, an acoustic model for parameter prediction, and a vocoder. The text analysis module first processes the text and extracts the linguistic features, such as phonemes, duration, and part-of-speech tags. The acoustic models, most notably hidden Markov models (Yamagishi et al. 2009) are trained with the paired linguistic features and acoustic features extracted from the speech. The vocoders synthesize speech from the predicted acoustic features. These approaches are superior to over previous TTS systems in terms of audio naturalness and controllability, while also requiring less recordings than concatenative synthesis. However, the generated speech has lower intelligibility due to artifacts, and the generated voice is perceived as robotic.

With the development of deep learning, neural network-based TTS has risen to prominence (Tan et al. 2021). We analyze some prominent neural TTS architectures in the next Subsection.

2.2.2 Neural Methods

WaveNet (Oord et al. 2016) can be regarded as the first modern neural TTS model, trained to directly generate audio waveforms from linguistic features. WaveNet is an autoregressive generative model for audio synthesis that uses dilated convolutions (Yu and Koltun 2015), with the predictive distribution for each audio sample conditioned on all previous ones. Its generated speech was more natural sounding than the state-of-the-art parametric and concatenative systems of the time. Following a similar autoregressive approach, Tacotron 2 (Shen et al. 2018) was proposed for realistic TTS, using spectrogram prediction followed a neural vocoder. The spectrogram prediction network is a recurrent, sequence-to-sequence model with encoder-decoder architecture that maps character sequences to mel spectrograms. The intermediate mel representation allows for separate training and is more compact than waveform samples, while being easier to train with squared loss due to phase invariance. Tacotron 2 still used WaveNet as the neural vocoder system.

Tachibana et al. (2018) were the first to use temporal convolutions for TTS, resulting in high parallelizability. The first fully parallel end-to-end architecture was proposed by Ma et al. (2020), and further improved inference speed in TTS. They tackled speech synthesis as a sequence-to-sequence mapping problem, using a modified U-Net architecture (Ronneberger et al. 2015). HiFi-GAN (Kong et al. 2020a) generates raw waveforms from the mel spectrogram, serving as an acoustic feature generator and a neural vocoder, respectively. It is a purely convolutional architecture that has demonstrated high-fidelity denoising and dereverberation in speech.

TTS is a multimodal alignment task, which many researchers attempt to guide with various attention methods. Badlani et al. (2022) introduce a diagonal static 2D prior distribution over the phoneme to acoustic feature mapping, thus accelerating learning. A similar prior is used in RAD-TTS (Shih et al. 2021), which is a flow-based parallel TTS model that resolves the output diversity issue by stochastically modeling the phoneme durations with a separate flow.

FastSpeech (Ren et al. 2019) and the improved FastSpeech 2 (Ren et al. 2020) tackle the one-to-many mapping problem in TTS by including a variance adaptor, offering control over the duration, energy, and pitch of the generated speech. Robust computational algorithms from traditional signal processing such as the short-time Fourier transform are incorporated in order to extract the speech parameters modeled by the variance adaptor. The model uses a transformer-based architecture and is thus able to predict the spectrogram significantly faster than older models. Ren et al. (2020) also propose FastSpeech 2s, a text-to-waveform model that does not

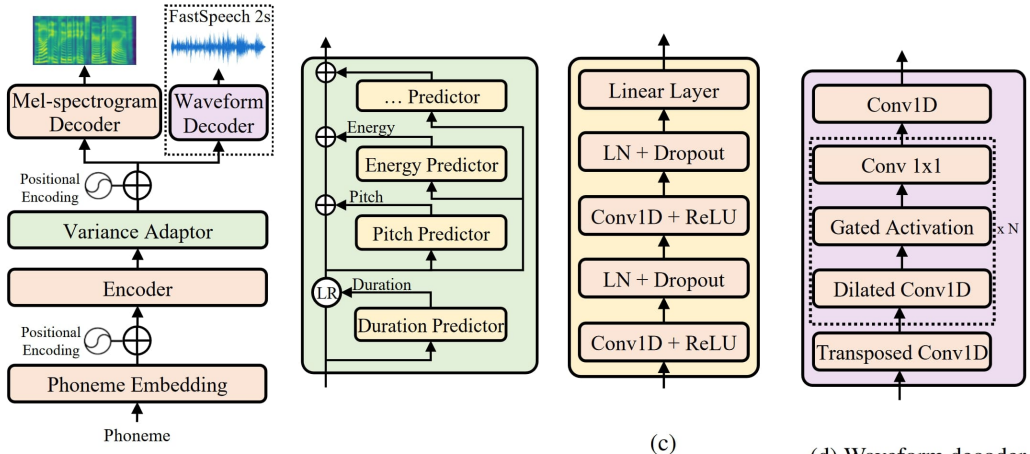


Figure 2.3: The FastSpeech 2 architecture and components, from Ren et al. (2020). Namely, from left to right:

- (a) FastSpeech 2 (text-to-spectrogram, left) and FastSpeech 2s (text-to-waveform, right) architectures. The encoder processes the phoneme embedding sequence, then the variance adaptor adds different variance information such as duration, pitch and energy into the hidden sequence. The FastSpeech 2 decoder predicts the mel spectrogram in parallel from the adapted hidden sequence into a mel spectrogram sequence. In contrast, FastSpeech 2s directly produces the speech waveform from the intermediate hidden representation.
- (b) The variance adaptor consists of duration, pitch, and energy predictors. They are trained using the corresponding ground truth values, so that they can predict those parameters during inference. The durations are used to expand the encoded phoneme sequence to match the length of the generated speech.
- (c) The internal architecture of a predictor submodule, using convolutions and layer normalization (Ba et al. 2016).
- (d) The waveform decoder of FastSpeech 2s, consisting of many convolutional layers that expand the encoded sequence in order to match the very large sequence length of the waveform.

use the intermediate spectrogram representation, achieving even faster inference. An overview of the architecture of FastSpeech 2 is presented in Figure 2.3.

Kim et al. (2020) propose Glow-TTS, a parallel flow-based generative model that uses a hard monotonic alignment search algorithm for robust TTS. Unlike FastSpeech, it doesn't require an external aligner, as the alignment search is incorporated inside the model, thus simplifying the training procedure. Wang et al. (2023) propose VALL-E, a model that uses residual vector quantization in order to treat TTS as a language modeling task. It has zero-shot and in-context learning capabilities, and can be used to synthesize high-quality personalized speech with only a 3-second recording of an unseen speaker as an acoustic prompt.

The unprecedented success of diffusion models quickly inspired works in speech synthesis. The first such attempts were by Chen et al. (2020b) and Kong et al. (2020b). The authors of the former achieved high-fidelity sampling in just six denoising steps, while the latter used fewer model parameters. Popov et al. (2021) propose Grad-TTS, a TTS system that includes an internal monotonic alignment search heavily based on Glow-TTS (Kim et al. 2020), but with a diffusion-based decoder.

Style Prompting

While the concept of style control has been widely explored in text or image generation, the introduction of explicit style controls in speech generation is still limited. Style guidance in TTS allows for more expressive speech and alleviates the one-to-many mapping problem. The most intuitive and interpretable method would be using natural language prompts. The first such attempt was by Kim et al. (2021b), who presented style tags as a novel style interface for expressive TTS. Style tags are short phrases or words, from which a linguistic embedding is extracted using a pretrained language model. Thus, the non-autoregressive TTS system can learn the relationship between linguistic embedding and style embedding space, and generalize well to unseen tags due to the generalization capabilities of the language model. Also, a reference style from an existing audio can be used as guidance. The audio-driven style encoder is trained to minimize the MSE between the linguistic embedding and the audio encoding vector.

PromptTTS (Guo et al. 2023) is a similar work that synthesizes expressive speech from a text prompt. The input prompt consists of a style prompt and a content prompt, formatted with a semicolon in between. The TTS system is a transformer-based encoder-decoder, infused with style information from the style encoder. A very recent approach on this subject that does not apply any constraints on the form of the natural language style prompts is InstructTTS (Yang et al. 2023). This method uses vector quantization in order to model acoustic features in discrete latent space, thus casting speech synthesis as a language modeling task. The vector-quantized acoustic features are predicted using a discrete diffusion model (Austin et al. 2021), while a cross-modal representation metric loss is also employed.

All the aforementioned works use variations of BERT (Devlin et al. 2018) to extract the style representations from the prompts. Regarding the available data, the three research teams used custom collected and annotated datasets. However, only PromptTTS's dataset (PromptSpeech) is both publicly available and is in English. Another interesting approach that does not accept text prompts but is purely conditioned on images is FACE-TTS (Lee et al. 2023). The proposed method extracts biometric features from a face image and uses them a condition to train a TTS model with diffusion. The model synthesizes speech whose pitch, rhythm and style matches the input face. A speaker feature binding loss is introduced in order to maintain speaker consistency between synthesized and reference speech.

Chapter 3

Face Modeling

Contents

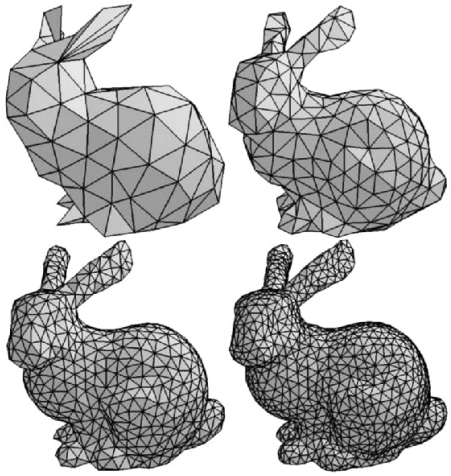


Figure 3.1: Triangular 3D geometric meshes of the same object, in different vertex resolutions. Figure from Novaković et al. (2017).

Preface

In the vast landscape of computer vision and graphics, one area that continually captivates researchers is face modeling, due to its numerous applications in entertainment, most notably in visual effects and computer generated imagery for games, films, or animations. More recent applications are in the areas of human-machine interaction, robotics, and virtual reality. Effective modeling of the human face is essential for realistic talking face generators. However, it is important to note that face models cover a much broader area of human endeavors. For instance, face identification through 3D models is used in biometric security systems or for forensic analysis. Specialized head modeling applications include medical imaging for craniofacial surgery, cognitive science and neuroscience.

3.1 Meshes

One of the most notable ways to represent 3-dimensional objects in graphics, is to encode them in *meshes*. Meshes are collections of N points in 3D space, called *vertices*, that are connected forming convex polygon *faces*. Triangular faces are the simplest and most widely-used way to represent the surface sampled by the vertices. In a triangular mesh, the vertices are organized in M triplets, producing the triangulated surface. Formally, the geometry of a 3D mesh is defined by the vertices table:

$$\mathbf{V} = [\vec{v}_1, ..., \vec{v}_N] \in \mathbb{R}^{3\times N} \tag{3.1}$$

where each element $\vec{v}_i = [x_i, y_i, z_i]^T \in \mathbb{R}^3$ describes the spatial coordinates of the i-th vertex. The topology is encoded in the triangle list of vertex triplets that encode the faces:

$$\mathbf{F} = [f_1, ..., f_M] \in \mathbb{R}^{3\times M} \tag{3.2}$$

where each element $f_i = (f_i^1, f_i^2, f_i^3)$ is a triplet of indices f_i^j that correspond to the vertices of the mesh. The mesh $\mathbf{M} = (\mathbf{V}, \mathbf{F})$ is a purely geometric representation, meaning that it doesn't involve any texture. An example of such a mesh is presented in Figure 3.1. Similarly, a textured mesh is represented by $\mathbf{M} = (\mathbf{V}, \mathbf{F}, \mathbf{C})$, with the texture $\mathbf{C}$ encoded as a per-vertex color vector of $\vec{c}_i = [r_i, g_i, b_i]^T \in \mathbb{R}^3$:

$$\mathbf{C} = [\vec{c}_1, ..., \vec{c}_N] \in \mathbb{R}^{3\times N} \tag{3.3}$$

The process of converting an abstract graphics object to an image is called *rasterization*. A 3D mesh can be rasterized in two steps. First, a rigid transformation relative to the camera shifts

and rotates the object. Then, a projection to the image plane each vertex to the pixel space. The projection may be perspective, weak perspective, or orthographic (Hartley and Zisserman 2003).

3.2 3D Morphable Models

One approach to 3D facial modeling would be to encode the facial geometry and color directly into a mesh. However, this does not take into account the high correlation of human faces, nor does it offer any statistical priors that can be sampled for generating new plausible faces. This is where morphable models come into play. The motivation behind them is to devise a way to express faces in a vector space. It is easy to see why, for instance, face images do not form a vector space. Assume two face images $\mathbf{I}_1, \mathbf{I}_2 \in \mathbb{I}$, where $\mathbb{I}$ is the field of all facial images. Averaging the two samples results in an image $\mathbf{I}$ with artifacts and blurred edges, which is not a plausible facial image. This is bound to happen no matter how well-centered the face is in the frame, since pixels from different positions on the two faces are averaged to generate the a new pixel value. Therefore, the need to separate the face shape and appearance arises, so that the faces can be linearly combined. This requires establishing establishing correspondence between the face representations.

3D Morphable Models (3DMMs) are statistical models of 3D shapes that separate shape from appearance variation, produced from datasets of 3D meshes. Typically, 3DMMs are used as prior probability distributions in computer graphics and vision, meaning that they can be sampled to render plausible shapes. The first 3DMM was pioneered by Blanz and Vetter (1999). The authors introduced a statistical model of human faces that could capture the variability of faces in a dataset.

The captured 3D meshes in such a dataset are highly correlated, since the characteristics of human faces are very similar. For instance, they all share the same rough geometry and attributes, which are similarly positioned on the head's surface. Factoring out all the shared characteristics, the true sources of variation in faces can be described with significantly fewer parameters. More formally, the data can be decorrelated and projected into compact feature spaces. Once built, the 3DMM serves two functions:

- It is a powerful prior on 3D face shape and texture, that can be leveraged in fitting algorithms to reconstruct accurate and complete 3D reconstructions of faces from data deficient sources, like in-the-wild 2D images, or noisy 3D depth scans.
- It provides a mechanism to encode any 3D face in a low dimensional feature space, a compact representation that makes tractable many 3D facial analysis problems.

In Chapter 5, we leverage the first function to reliably obtain accurate 3D reconstructions of subjects while speaking, in order to train our audiovisual speech model. We also make use of the 3DMM's low dimensionality in order to efficiently predict 3D talking heads from a text encoding.

3.2.1 Construction

3DMMs are learned from high-quality 3D scans that must be aligned and brought to a common reference, where the geometry and topology are consistent across all meshes, thus achieving *dense correspondence*. Sparse correspondences have also been used Egger et al. (2020), extracted via landmarks or local image descriptors like Scale-Invariant Feature Transform (SIFT) features (Lowe 2004).

In order to establish dense correspondence, all the training meshes need to be reparameterized into a representation where each of them has the same number of vertices, and all of them share the same triangulation. Dense correspondence algorithms rely on a template mesh onto which the target meshes need to be mapped to. This mapping can be done in two ways:

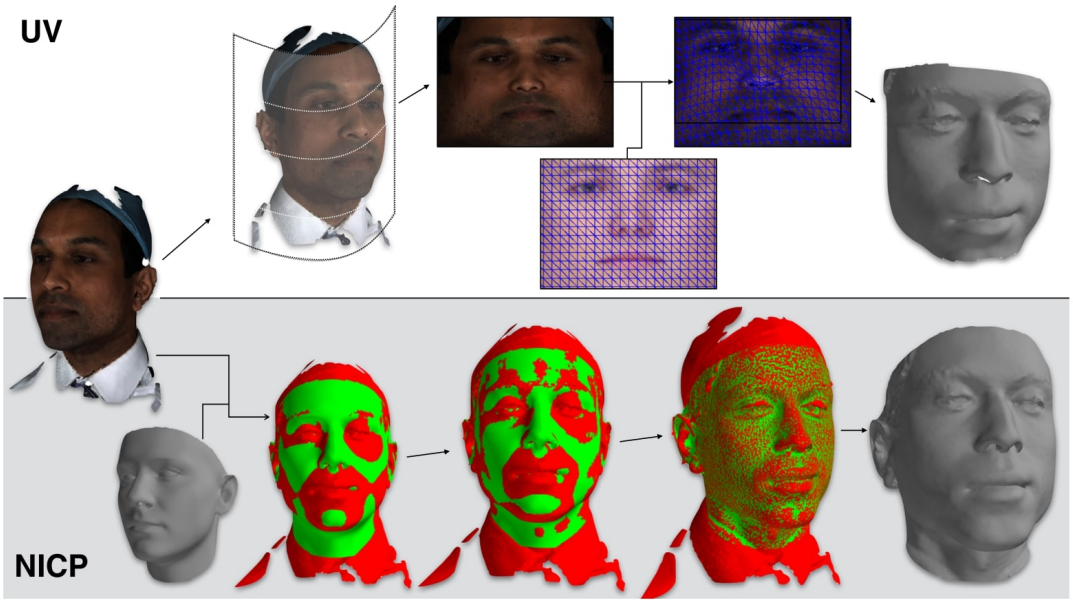


Figure 3.2: Establishing facial mesh correspondence using either a UV projection (top) or NICP (bottom). In the former, the cylindrical projection of the 3D mesh is deformed in 2D to match the vertices of the 2D template. In the latter, the template 3D surface is manipulated with in order to match the sample. Figure from Booth et al. (2018).

1. Using an intermediate 2D *UV* space, can be thought of as a flattened atlas where the 3D face surface has been projected. By assuming accurate representation of the 3D facial structure into the UV space, the problem is reduced to an image alignment task. In Blanz and Vetter (1999), the authors used optical flow estimation in order to align UV maps.

2. Using Non-rigid Iterative Closest Point (NICP), an algorithm that estimates affine transformations for each vertex of the template, so that the transformed vertices form a mesh that is as close as possible to the target.

The collection of meshes in dense correspondence are statistically analyzed, typically with *Principal Component Analysis* (PCA), generating a 3D deformable model as a linear basis of shapes. After performing PCA, linear models are obtained for each source of variation. The resulting shapes and textures are described by vectors of $3N$ elements. For instance, the shape mesh $\mathbf{S}$ is flattened to a vector $\vec{S}$, consisting of the mean shape $\bar{\vec{S}}$ plus a linear combination of *eigenshapes*:

$$\vec{S} = \bar{\vec{S}} + \sum_{i=1}^{d_{\text{shape}}} \beta_i \vec{U}_i = \bar{\vec{S}} + \mathbf{U}\vec{\beta} \tag{3.4}$$

where $\mathbf{U} = [\vec{U}_1, ..., \vec{U}_{d_{\text{shape}}}] \in \mathbb{R}^{3N \times d_{\text{shape}}}$ is the orthonormal basis matrix whose columns contain the shape eigenvectors $\vec{U}^i$ and $\vec{\beta} \in \mathbb{R}^{d_{\text{shape}}}$ is the shape vector. Thus, any 3D face mesh can be efficiently compacted into a vector representation $\vec{\beta}$ that contains only a few parameters. The eigenvectors do not coincide with attributes that humans would use to describe a face, since they are extracted via mathematical analysis. It is also interesting to note that the components of the shape vector tend to be small. As the coefficients grow, the face becomes more characteristic, eventually leading to a caricature. The magnitude of the shape vector is indicative of failure cases, if it exceeds a certain threshold (Booth et al. 2018). This implies that each coefficient vector needs an assigned probability of describing a realistic face. The probability distribution is typically assumed to be Gaussian with a block diagonal matrix, which assumes that shape and texture are decorrelated.

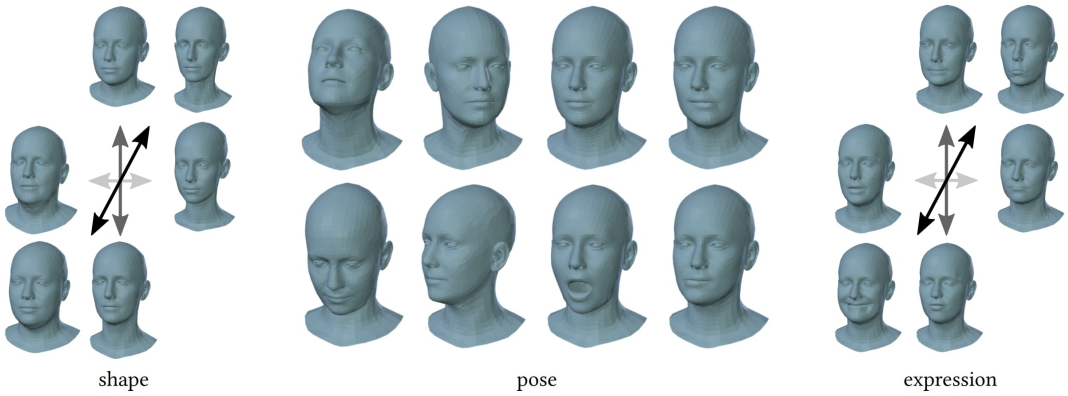


Figure 3.3: Separation of sources of variation in the FLAME model. The figure shows the activation of the first three principal components of shape and expression at ±3 standard deviations, as well as pose parameters actuating neck and jaw joints. Figure from (Li et al. 2017).

Facial expression can be incorporated to the model by adding the expression variation:

$$\vec{S} = \vec{\bar{S}} + \mathbf{U}\vec{\beta} + \mathbf{V}\vec{\psi} \tag{3.5}$$

with $\mathbf{V} \in \mathbb{R}^{3N \times d_{\text{exp}}}$ being the orthonormal basis matrix of expression eigenvectors, and $\vec{\psi} \in \mathbb{R}^{d_{\text{exp}}}$ being the expression coefficients vector.

Similarly, a mesh's texture can be expressed as the mean texture $\vec{T}$ with the addition of weighted orthoromal texture components, extracted as described above:

$$\vec{T} = \vec{\bar{T}} + \mathbf{C}\vec{c} \tag{3.6}$$

with $\mathbf{C} \in \mathbb{R}^{3N \times d_{\text{tex}}}$ the orthonormal basis matrix whose columns contain texture eigenvector, and $\vec{c} \in \mathbb{R}^{d_{\text{tex}}}$ the vector of texture coefficients.

3.2.2 Prominent Models

Large-Scale Facial Model

Booth et al. (2018) propose the Large-Scale Facial Model (LSFM), which is constructed from a large and diverse database of face scans. The authors establish correspondence using the NICP algorithm, focusing on the anatomical structure of the face. Thus, they avoid any alignment based on "skin-deep" facial features like eyebrows. They find that the manifold of plausible faces is clustered by demographics and exhibits an age-related structure.

FLAME

In this Thesis, we make use of the FLAME 3DMM (Li et al. 2017), a statistical face model that separates shape, expression, head pose, and texture. It is trained on sequences of 3D scans and is able to capture correlations across the face, and capturing realistic *blendshapes*, which are approximate semantic parameterizations of facial expression. The FLAME 3DMM is described as a function

$$\mathcal{M}(\vec{\beta}, \vec{\theta}, \vec{\psi}) \rightarrow (\mathbf{V}, \mathbf{F}) \tag{3.7}$$

that maps the vectors for shape ($\vec{\beta}$), pose ($\vec{\theta}$) and expression ($\vec{\psi}$) to a 3D displacements of the template mesh. See Figure 3.3 for a visualization of the different sources of variation in the FLAME model.

Shape and expression are modeled linearly, using orthonormal eigenvectors. However, pose is a non-linear function $R(\vec{\theta})$ that maps the pose vector to a vector of concatenated elements of

rotation matrices, which are the reason for the mapping's non-linear nature. Then, the pose blendshape is defined as a linear combination of pose blendshape components, with the offsets from the rest pose as weights. The integration of pose is what allows FLAME to achieve better human expressions, by modeling four joints with 3D rotations: the neck, the jaw, and the two eyeballs. A non-linear function is used to map the pose vector to the elements of each rotation matrix. The computation of joint locations from mesh vertices is done by a learned sparse matrix. The model parameters are trained with the objective of minimizing the 3D reconstruction error.

Weighting is used for improved accuracy, especially in regions that tend to be noisy in the scans, such as the hair, the eyeballs and the back of the head. Coupling weights prevent the vertices from leaving the model space, while Laplacian weights add smoothness while allowing tangential motion to be captured.

3.3 Facial Reconstruction Methods

In this Thesis, since we are interested in audiovisual speech synthesis, we use 3DMMs as a tool for face modeling. However, we cannot expect to train our models with audiovisual 3D scans of high quality audio and transcriptions of the uttered speech, since they are expensive to capture and cannot generalize to other subjects outside the dataset. We need to leverage the already existing video datasets, which is the motivation behind using 3DMMs as a prior for in-the-wild accurate head reconstruction in 3D space.

Originally, Blanz and Vetter (1999) reconstructed human faces from images by estimating the 3DMM parameters for shape and appearance, as well as camera and illumination parameters. They did this by performing stochastic gradient descent for minimizing the $L2$ distance between the original image, and the image with the rendered 3DMM on top. However, estimating a 3D shape from an image is an ill-posed problem and may yield non-face solutions. In order to deal with this issue, they restricted the shape and texture vectors to the vector space spanned by the 3D database.

Since then, there have been many approaches to facial reconstruction from images (Zollhöfer et al. 2018). A natural limitation that audiovisual speech enforces is that it needs to be expressive in order to convey meaning. Thus, in the next Subsections, we analyze some recent models.

3.3.1 Detailed and Emotional Reconstruction

DECA

DECA, a model for Detailed Expression Capture and Animation (Feng et al. 2021) is a 3D reconstruction method with a FLAME prior that emphasizes fine-grained details in expression. It is trained on in-the-wild images, achieving better robustness and more realism in animation than previous approaches. DECA regresses 3D face shape and animatable details like pores and wrinkles that are specific to an individual subject, but change with expression, all from from a single image of the subject.

The method reconstructs face images in two scales, a coarse one and a detailed one. The coarse reconstruction employs a ResNet50 (He et al. 2016) encoder followed by a fully connected layer in order to predict FLAME and environment parameters for an input image. Formally, given an image $\mathbf{I}$, the coarse encoder $\mathcal{E}_c$ outputs the parameters:

$$\mathcal{E}_c(\mathbf{I}) \rightarrow (\vec{\beta}, \vec{\theta}, \vec{\psi}, \vec{l}, \vec{c}, \mathbf{A}) \tag{3.8}$$

with $\vec{l}$, $\vec{c}$, and $\mathbf{A}$ being the lighting parameters, camera parameters, and albedo image, respectively. On the other hand, the detail reconstruction extends the coarse geometry by superimposing a higher-frequency UV displacement map. The 2D displacement map is decoded from a detail code predicted by an encoder $\mathcal{E}_d$ similar to the coarse one:

$$\mathcal{E}_d(\mathbf{I}) \rightarrow \vec{\delta} \tag{3.9}$$

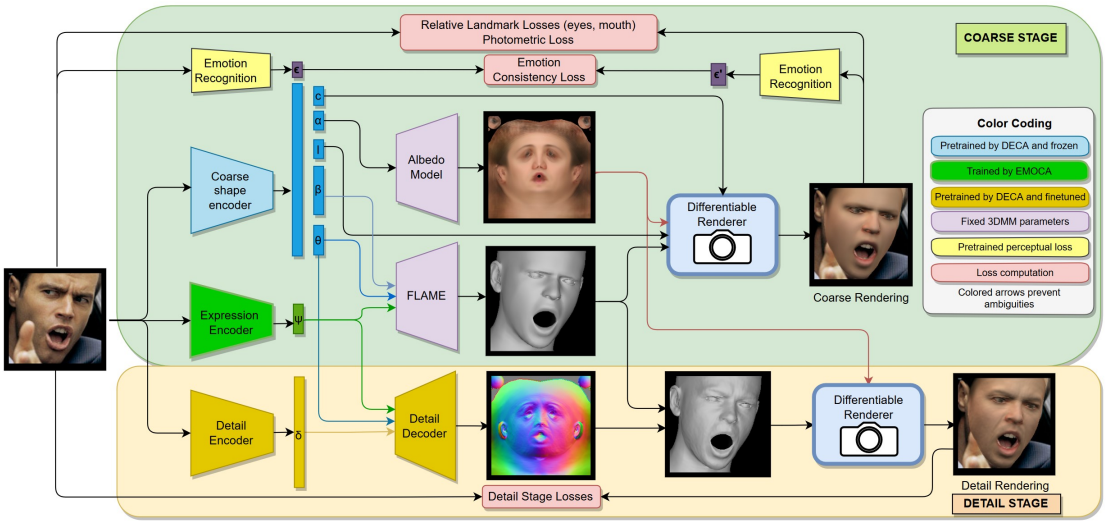


Figure 3.4: Architecture and training procedure of EMOCA, an extension of DECA. The input image is fed to a frozen DECA coarse shape encoder and a trainable expression encoder. A textured 3D mesh is rendered by a differentiable renderer with the regressed camera and spherical harmonics lighting. A novel emotion consistency loss penalizes the difference between the emotion features of the input and the rendered coarse shape, after passing both images through a fixed emotion recognition network. By factoring out the separate expression encoder and the emotion recognition network, the architecture effectively becomes identical to DECA. Figure from Daněček et al. (2022).

In order to disentangle person-specific details like skin wrinkles, from expression-dependent ones, a detail consistency loss is employed. This loss is devised as the distance between two reconstructions, one using the detail code of the original image, and the other using the detail code of another image of the same subject. The intuition behind this is that the detail codes of the two images should yield the same rendering. This allows the detail vector to capture all person-specific higher-frequency attributes and leave out the expression-dependent details which vary from image to image.

EMOCA

Similarly, EMOCA is a model for Emotional Capture and Animation (Daněček et al. 2022) that builds on DECA, aiming to better capture the human expressions associated with emotions. The authors augment DECA's architecture by using a separate encoder to estimate the expression vector and using out-of-the box DECA estimations for the remaining FLAME and scene parameters. They also introduce a deep perceptual emotion consistency loss, using a pretrained emotion regressor to find the distance between the reconstructed 3D expression, and the input image in the emotional feature space. This is expressed as:

$$\mathcal{L}_{emo} = ||\vec{\epsilon}_I - \vec{\epsilon}_R||_2 \tag{3.10}$$

where $\vec{\epsilon}_I$ and $\vec{\epsilon}_R$ are the emotion feature vectors of the original and rendered images, respectively, predicted by the pretrained emotion recognition network. We present EMOCA's architecture in Figure 3.4.

3.3.2 Speech-Informed Perceptual Reconstruction

In this Subsection we present SPECTRE (Filntisis et al. 2023), a method for perceptual 3D reconstruction of human face videos focusing on lip articulation, without the need for text transcriptions of the corresponding speech. A *lipreading* loss is devised to add more realism to the mouth region, improving the perceived quality of the utterance.

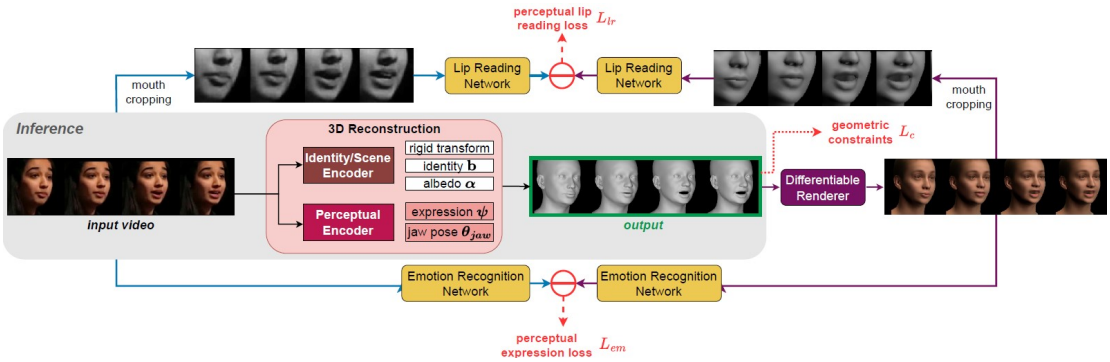


Figure 3.5: Architecture and training procedure of SPECTRE. The input video is fed into the 3D reconstruction component, where a fixed encoder detects the scene and FLAME parameters. Then, a mouth encoder predicts the refined facial expression parameters and jaw pose, while a differentiable renderer renders the predicted 3D shape. The mouth area is cropped in both the input and rendered image sequences and a lip reader is applied on both in order to estimate the perceptual lip reading loss between them. The same is done for the facial expression recognizer, in order to estimate the perceptual expression loss. Figure from Filntisis et al. (2023).

The pipeline is similar to EMOCA, but the focus is shifted on the realism of lip reconstruction. SPECTRE uses a perceptual encoder to estimate the mouth parameters, which consist of the expression and jaw pose. This encoder is built with a MobileNetV2 (Sandler et al. 2018) backbone followed by a temporal convolution, in order to capture the temporal dynamics of mouth movements and facial expressions. The MobileNetV2 encoder was consciously chosen in order to mitigate the heavy computational requirements of the ResNet50 encoder used by EMOCA. A perceptual expression loss is applied between emotional feature vectors of the input video and the reconstructed 3D mesh.

Furthermore, a pretrained lipreading network on the LRS3 dataset (Afouras et al. 2018b) extracts speech-informed feature vectors which are used to calculate the lipreading loss as the cosine distance of feature vectors from the original and the reconstructed video.

Geometric constraints are also applied on the magnitude of the mouth parameters using $L2$ norm, to avoid any artifacts that perceptual losses may create due to the domain mismatch between the original and the rendered images. We present SPECTRE's architecture in Figure 3.5.

3.3.3 Other Approaches

An issue with the modeling of human faces, already pointed out by Booth et al. (2018), is the metric accuracy of the reconstructed face. A wide area of applications in medicine and augmented or virtual reality rely on the metrically correct prediction of the subject's face. MICA (Zielonka et al. 2022) tackles this problem by training a face shape estimator in a supervised fashion, using annotated medium-scale datasets and data from a large-scale 2D image database, processed by a pretrained face estimation network.

Another approach that doesn't rely on latent subspaces is the use of *Convolutional Neural Networks* (CNNs). In Jackson et al. (2017), a CNN performs direct regression of a volumetric representation of the 3D facial geometry from a single 2D image. Later, in Abrevaya et al. (2018), they employ an autoencoder architecture, using a CNN encoder that operates on depth images. Ranjan et al. (2018) use spectral convolutions on the mesh, represented as a graph. In Zhou et al. (2020) they advance this idea by using a fully convolutional mesh autoencoder for arbitrary registered mesh data, outperforming previous state-of-the-art methods.

Chatziagapi and Samaras (2023) incorporate the audio modality into 3DMMs by proposing AVFace, a method for 4D facial reconstruction using audiovisual information. In their model, a coarse stage estimates the per-frame FLAME parameters using AV features. Then, an implicit

representation of the lip shape conditioned on speech is learned. Finally, a fine stage recovers facial geometric details guided by pseudo-ground truth face normals. Due to the multimodal features, AVFace is robust in cases when either modality is insufficient, for instance in cases where face is occluded.

Chapter 4

Talking Face Generation

Contents

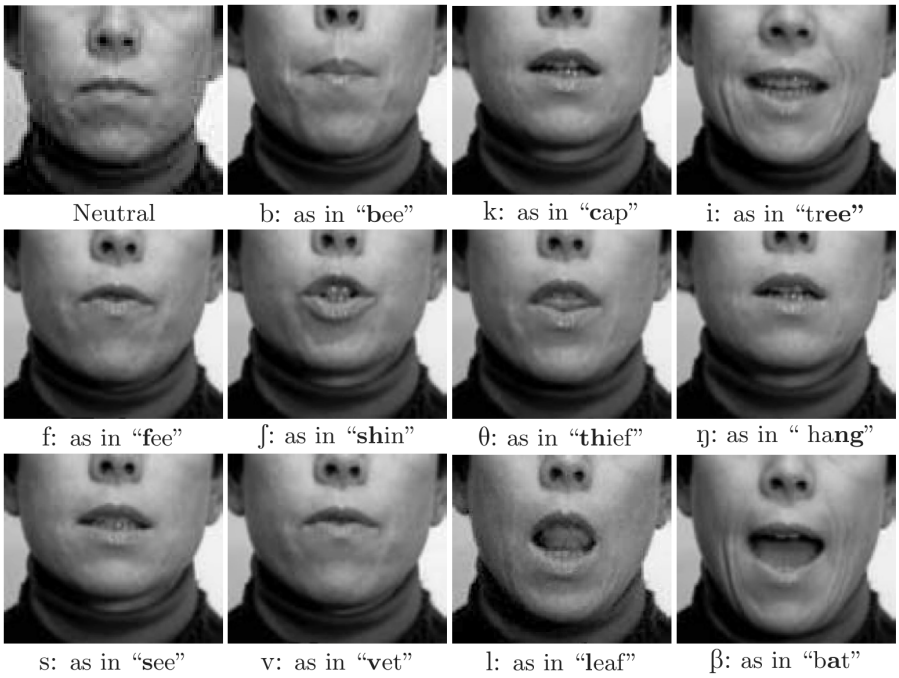


Figure 4.1: A few viseme examples. This Figure is from Parke and Waters (2008) and uses the International Phonetic Alphabet instead of the ARPAbet notation which we presented in Table 2.1. Nevertheless, it showcases the variety of visemes when it comes to the relative positions of the lips, tongue, and teeth during articulation.

4.1 Introduction

Talking face generation aims at synthesizing videos of talking humans, with consistent and synced audio and visual streams. Its applications are numerous, including virtual assistants, educational programs, accessibility tools, and human-machine interfaces (Toshpulatov et al. 2023). Other notable applications lie in the field of entertainment, where it is used for avatar creation in movies and video games, or re-dubbing movies in foreign languages. Furthermore, it has been used for teleconferencing applications, such as compressing video calls (Agarwal et al. 2022).

In order to produce realistic results, face representations are usually adopted in order to guide the generation process. Those representations may be photorealistic, 3D mesh-based, or parametric, by using an intermediate representation. Examples of parametric modeling approaches are the image-based active appearance models (Cootes et al. 2001), or the prominent 3DMMs presented in Section 3.2. 3D methods offer increased accuracy and realism, since they are robust to pose and lighting changes. Unlike traditional 2D images, which capture only surface appearance, 3D models encapsulate the underlying geometry of facial structures, including shape, depth, and spatial relationships. Also, by treating faces as 3D objects, we gain the ability to simulate realistic lighting effects and intricate facial movements. Nevertheless, in pure 3D-based works, the absence of teeth and tongue is important, since they play a large role in the realism of specific types of *visemes*. Visemes are visual analogs to phonemes, but the phoneme to viseme mapping is many-to-one, implying that the uttering of some phonemes is visually the same. Such examples include the phonemes `B`, `M`, and `P`. We present a few viseme examples in Figure 4.1.

Human perception has evolved to be very sensitive to changes in facial characteristics, and can effortlessly distinguish between similar faces or recognize whether a face looks unnatural (Mori et al. 2012). Thus, face models need to be as accurate as possible. Apart from realistic head modeling, talking face generation needs to produce videos with coherent and consistent audio

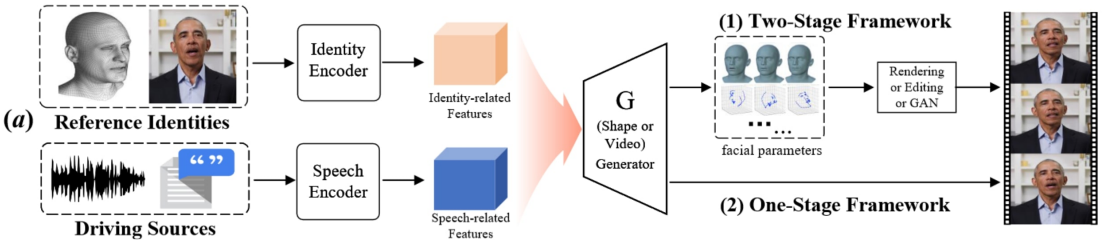


Figure 4.2: A general depiction of a talking face generation pipeline. The inputs are a reference identity for the talking subject, as well as a driving source for the utterance, be it audio or text. The synthetic video result can either use an intermediate parametric representation of human faces, or be directly computed from the input features. Figure from Sheng et al. (2022).

and visual streams. The generation can either be audio-driven, meaning that a talking face video is synthesized to match an input audio clip, or text-driven, where both audio and video streams have to generated from a piece of natural language text. In this Diploma Thesis, we developed a text-driven model that uses an intermediate 3D representation. A general overview of generative talking face systems is presented in Figure 4.2.

4.2 Audio-driven Methods

Audio-driven audiovisual speech synthesis aims to produce a realistic visual output of a talking head that is accurately synced with the input audio. Most cutting-edge approaches are audio-driven, since one can argue that they are more manageable than text-driven frameworks in terms of data requirements, while still covering a wide array of applications (Sheng et al. 2022). Nevertheless, they need to employ an external TTS method in order to be able to create fully customized talking face videos, with arbitrary utterances.

Neural Voice Puppetry, by Thies et al. (2020) is a deepfake pipeline that generates a photorealistic output video of a target person, in sync with some source audio input. The system uses a pretrained speech feature extractor (Hannun et al. 2014), a per-frame blendshape generator whose output spans a latent expression space, and a neural renderer. The expression interpolator and the renderer have to be trained on each specific output target, as shown in Figure 4.3. Fan et al. (2022) proposed a model that generates a sequence of 3D meshes that match an audio input, however they tackled the long-term context limitations of recurrent models by using the transformer architecture. CodeTalker (Xing et al. 2023) is a similar model that achieves the current state-of-the-art in audio-driven 3D mesh generation.

Accurate, human-like 3D reconstruction of a talking face is an open problem. A video of a talking person contains rich dynamic information about pose and expression, especially in the mouth region, which needs to be explicitly modeled. Attempts like VOCA (Cudeiro et al. 2019) extract speech features from the well established speech recognition RNN DeepSpeech (Hannun et al. 2014) and use time convolutions to convert them to 3D mesh displacements with an encoder-decoder architecture. Bao et al. (2023) take this approach one step further by using audio features, LSTMs and phoneme guidance to produce viseme curves, suitable for facial animation. Their method supports multilingual speech inputs and generalizes well to unseen speakers.

Wav2Lip (Prajwal et al. 2020) generates synchronized lip movements from an input speech signal. The model consists of a visual feature extractor, a lip motion generator, and a temporal alignment module. The visual feature extractor encodes the visual features of the input video frames, while the lip motion generator produces lip movements corresponding to the input audio signal. These components are trained jointly to optimize the synchronization between the generated lip movements and the audio. To achieve accurate temporal alignment between the audio and visual modalities, the model employs a temporal alignment module that leverages dynamic time warping techniques. This module helps mitigate discrepancies in the temporal

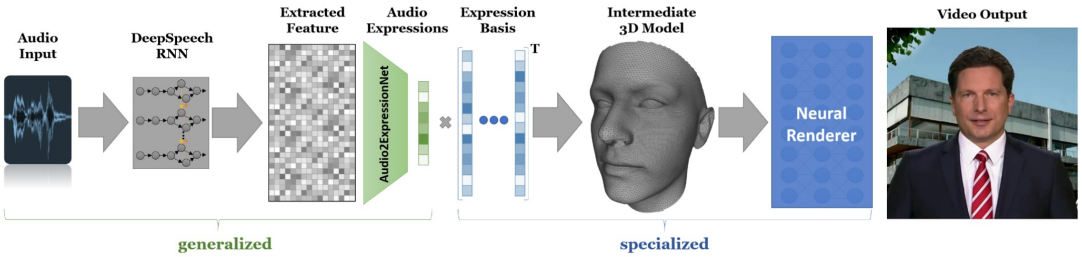


Figure 4.3: The architecture of Neural Voice Puppetry. Given an audio signal, the DeepSpeech RNN (Hannun et al. 2014) predicts speech features that are fed into a generalized expression prediction network, which predicts coefficients that drive a person-specific expression blendshape basis. The target face is rendered with the new expressions using a light-weight rendering network based on the U-Net architecture. A similar pipeline is common across audio-driven works. Note that this work combines a generalized and a person-specific stage. Figure from Thies et al. (2020).

dynamics of speech and lip movements, resulting in more natural and coherent output.

Other few-shot lip-syncing methods include SadTalker (Zhang et al. 2022c) and VideoReTalking (Cheng et al. 2022). The former extends Wav2Lip's architecture with a 3DMM face representation, achieving better control of facial expressions. The latter achieves state-of-the-art results in audio-driven photorealistic talking head generation. Its architecture is presented in Figure 4.4.

4.2.1 GAN-based Models

Neural Emotion Director, from Paraperas Papantoniou et al. (2022), is a deepfake generator that can manipulate in-the-wild videos conditioned on either an emotional label or a reference video, while preserving the speech-related mouth movements. The emotion translation problem is mapped from the image space to the space of 3D model parameters using DECA (Feng et al. 2021). The expressions are altered with a recurrent model trained in an adversarial manner, then a neural renderer is used to create the new cropped face video. Several recent methods synthesize photorealistic facial videos using conditional GANs. Methods like Deep Video Portraits (Kim et al. 2018), Head2HeadKim et al. (2018), Doukas et al. (2021a), and Doukas et al. (2021c) use conditional GANs to render the target subject under the given conditions (expressions, pose, eye-gaze).

However, these methods need a driving video of an actor's face and do not offer any semantic control over the generated video. This is partly overcome by methods that offer control in terms of facial expressions (Tripathy et al. 2020; Tripathy et al. 2021; Groth et al. 2020; Solanki and Roussos 2021), without however having any control or constraints on the speech-related facial motions. Kim et al. (2019) presented a style-preserving solution to film dubbing, where the expression parameters of the dubber pass through a style-translation network before driving the performance of the foreign actor. Their method preserves the dubber's speech, but can only translate between a pair of speaking-styles (dubber-to-actor).

4.2.2 Diffusion-based Models

In Diffused Heads, from Stypułkowski et al. (2023), realistic talking head videos are produced with one identity frame and a speech recording, employing a U-Net architecture with several conditioning mechanisms. The authors concatenate the identity frame along with the two previous video frames and the current noisy frame channel-wise, and perform the denoising process, conditioned on a speech window. The conditioning is done with group normalization (Wu and He 2018), injecting information from a pretrained audio encoder. For improved realism, they incorporate a lip-sync loss defined as the MSE in the cropped mouth region, in addition to the

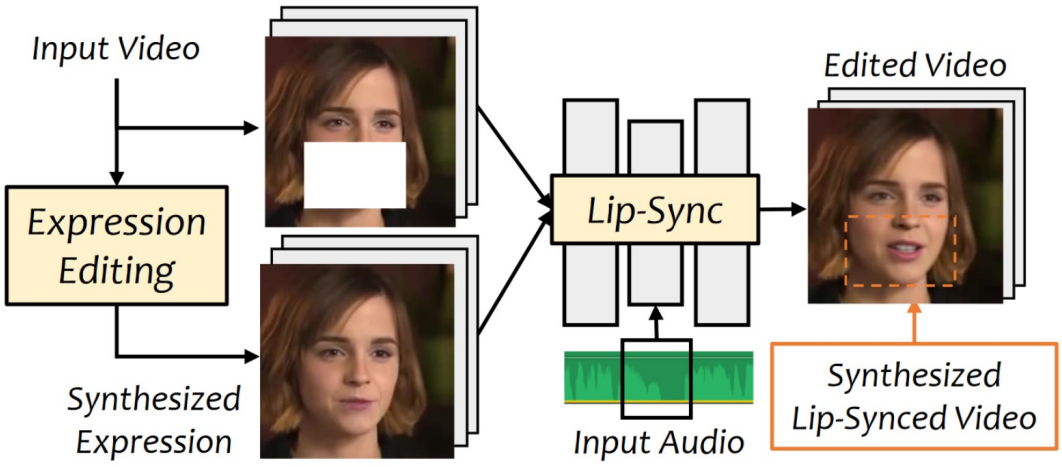


Figure 4.4: VideoReTalking alters the speaker's expression using a semantic expression editing network, resulting in a video with the neutral expression, which will serve as the reference for lip syncing. This video, together with the given audio, is then fed into the lip-sync network to generate the lip-syncing video. Finally, the photorealism of synthesized faces is improved through an identity-aware face enhancement network and post-processing. Figure from Cheng et al. (2022).

simple diffusion loss. A similar, concurrent work is DiffTalk (Shen et al. 2023). The authors employ a U-Net with cross attention for the video generation. They also downsample the frames for the diffusion process, while Diffused Heads use full resolution end-to-end. Apart from audio conditioning, they concatenate a facial landmark embedding along the axis of the audio conditioning vector.

Diffusion-Autoencoder-Talker, by Du et al. (2023), is another diffusion-based work that utilizes a different approach. An image encoder is trained to extract a latent representation of talking head frames, with a DDIM decoder to reconstruct them. Then, a convolutional transformer (Gulati et al. 2020) is trained to extract the same latent representation from a corresponding speech window. Inference is performed by denoising the same initial image for each frame with the DDIM decoder, conditioned on the audio window from the speech input.

Yu et al. (2022) perform one-shot audio-driven talking head generation with diverse facial motions, by disentangling the lip and non-lip facial representations. They use a diffusion prior trained to model the mapping between audio and non-lip features, thus enabling diverse sampling. This mitigates the lack of expressiveness and head motions often encountered in synthesized talking heads. Hwang et al. (2023) propose a model that disentangles head motion from facial expressions, using a geometric transformation as a bottleneck for the head motion. The framework synthesizes a talking head video based on a source video, a reference head motion and a driving audio. A motion-aware encoder captures the input talking head video, along with the motion of a head driving video, using either optical flow or neural features. Then, the intermediate representation is manipulated by convolution layers whose weights are modulated with audio features from the driving audio.

In contrast, Bigioi et al. (2023) explicitly edit an input video to incorporate a target audio. They use a diffusion model with U-Net architecture, conditioned on the masked video frame and corresponding speech features. Yao et al. (2021) present a text-based tool for iterative editing, where users can edit the wording of the speech, further refine mouth motions, insert mouth gestures and change the performance style. Their method uses a fast phoneme search algorithm that can identify phonemes from the source video to match a desired edit. They can also transfer the mouth motions of the source actor to the target actor with neural retargeting.

4.2.3 Emotional Talking Faces

Audio-driven emotional talking head generation has been also tackled by Ji et al. (2022), with a self-supervised model that generates emotional talking heads using an input audio and a reference emotional video. A recurrent network models neutral expressions, while a dynamic part predicts keypoint displacements from the input audio. EmoTalk (Peng et al. 2023) explicitly disentangles emotion from audio content. It uses an emotion disentangling encoder trained with cross-reconstruction loss, using pairs of the same content with different emotion and vice-versa. Then, a blendshape decoder predicts the head coefficients which are rendered using FLAME (Li et al. 2017).

It is important to note that audio-driven emotional talking head generation lacks control of the input audio, which is naturally correlated with the emotion. Emotional manipulation is still an active area of research, facilitated by large datasets like MEAD (Wang et al. 2020), a multi-view audiovisual dataset with diverse subjects and emotion annotations, suitable for emotional talking-face generation. A text-annotated version of the MEAD dataset is used in TalkCLIP (Ma et al. 2023), a framework for audio-driven talking head generation with explicit style control either from a reference video or a natural language prompt.

4.3 Text-driven Methods

Text-driven audiovisual speech synthesis involves the generation of both audio and visual outputs for a talking face, based on a text input. It is thus more challenging than the audio-driven approach, but has broader applications. One of the first attempts in the field was by Ezzat and Poggio (2000), who mapped text to visemes, creating a photorealistic viseme sequence from recorded frames. The transition between frames was smoothed using optical flow methods. Another approach by Filntisis et al. (2017a) uses active appearance models for the generated face image textures, and hidden Markov models for the generation of audiovisual features, focusing on the photorealism and the expressiveness of the output video.

Filntisis et al. (2017b) perform a comparison of deep learning, HMMs, and concatenative models for audiovisual speech synthesis with active appearance models, focusing on the expressiveness of the generated speech. They conclude that deep learning approaches outperform the traditional ones. They also show that realistic interpolation of emotions is possible with HMMs, allowing for both intermediate emotions and lower intensity of a particular emotion, by interpolating with a neutral speaking style.

4.3.1 Cascaded or Unimodal Methods

A common approach for text-driven models is to use a cascaded architecture of a TTS module connected in series with audio-driven talking face generator. Such models include ObamaNet (Kumar et al. 2017) and AnyoneNet (Wang et al. 2022). The former can be fine-tuned to one speaker identity, while the latter can perform one-shot talking head generation given an input portrait, while the generated audio matches the facial attributes of the person. A similar approach is proposed by Obradović et al. (2022), who use Wav2Lip (Prajwal et al. 2020) on top of a TTS framework. Zhang et al. (2022b) use TTS to generate speech from an input text, apply forced alignment to obtain phoneme timestamps, and lookup phoneme poses in a phoneme-pose dictionary. The poses are then interpolated and rendered with a GAN. A similar pipeline is used by Song et al. (2022) for multilingual talking face generation, and Ye et al. (2023) who propose a talking face generator that follows a voice cloning TTS module.

While the abundance of TTS and audio-driven models make this approach quite promising, the information flow is not as efficient. The intermediate representations that are extracted in the TTS system are then interpolated with redundant information in order to synthesize realistic speech, which then has to be compactly encoded once more. Furthermore, this approach doesn't consider the correlation of audio and visual streams in human speech, which is inherently bimodal.

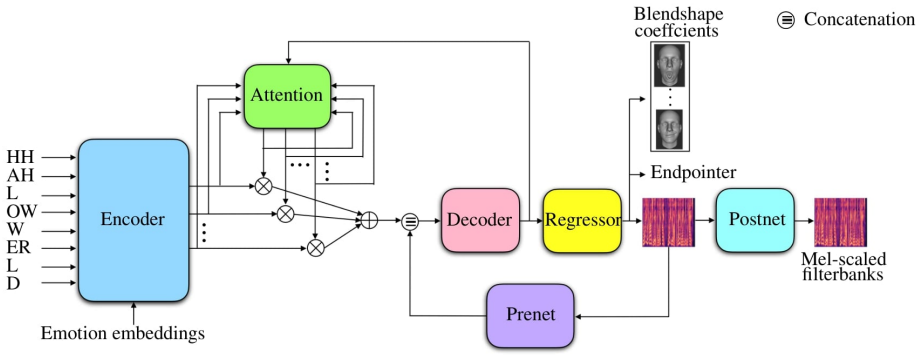


Figure 4.5: The AVTacotron2 model for autoregressive text-driven audiovisual speech synthesis. An autoregressive encoder-decoder architecture generates audiovisual representations which are jointly mapped to a spectrogram and 3D face parameters. Emotional embeddings offer explicit control using predefined emotion labels. Figure from Abdelaziz et al. (2021).

A few research works focus on text-to-lip generation. For instance, Liu et al. (2022) propose a parallel model for fast generation of cropped lip image sequences from an input text. Using structural similarity index loss and adversarial learning, they significantly improve the sharpness and perceptual quality of generated lip frames. Write-a-Speaker (Li et al. 2021) generates photorealistic talking head videos with realistic facial expressions and head motions in accordance with speech context and rhythm. A speaker-independent stage includes three disentangled networks that use the input text to generate animation parameters of the mouth, upper face, and head, drives the speaker-specific stage, which synthesizes videos tailored for different individuals. The three disentangled networks are trained with specific losses for each task, including adversarial loss for improved diversity. However, by not including the audio modality, the correlation of speech with lips is disregarded, which doesn't make text-to-lip models flexible enough to be used alongside a TTS system. Regarding 3D-based approaches, most authors note that the non-parameterization of the mouth interior naturally leads to less lifelike viseme results. Medina et al. (2022) directly address this limitation by providing an inner mouth dataset and model. Another interesting work that incorporates text is a tool for iterative editing is presented by Yao et al. (2021). Its users can edit the wording of the speech, further refine mouth motions, insert mouth gestures and change the performance style.

4.3.2 Audiovisual Methods

A true text-driven neural model for audiovisual speech synthesis was proposed by Abdelaziz et al. (2021), who expanded the Tacotron 2 TTS framework (Shen et al. 2018) to include a visual modality, and proposed the AVTacotron2 model. It produces emotional speech by using emotion embeddings to encode the required prosody. The corresponding blendshapes and spectrograms are generated in an autoregressive manner, conditioned on the text's phoneme encodings. A diagram depicting the system's architecture is depicted in Figure 4.5.

Similarly, DurIAN (Yu et al. 2019) adapts the WaveRNN model (Kalchbrenner et al. 2018) to generate facial animation parameters. UniFLG (Mitsui et al. 2023) learns a joint representation of text and audio, thus enabling both text-driven and audio-driven synthesis. Either way, the model's output is a sequence of facial landmarks, as well as an audio clip in text-driven inference. The facial modeling using landmarks, however, is not a suitable representation for high-detail lip articulation and cannot generalize to new faces. The inference process of UniFLG is depicted in Figure 4.6.

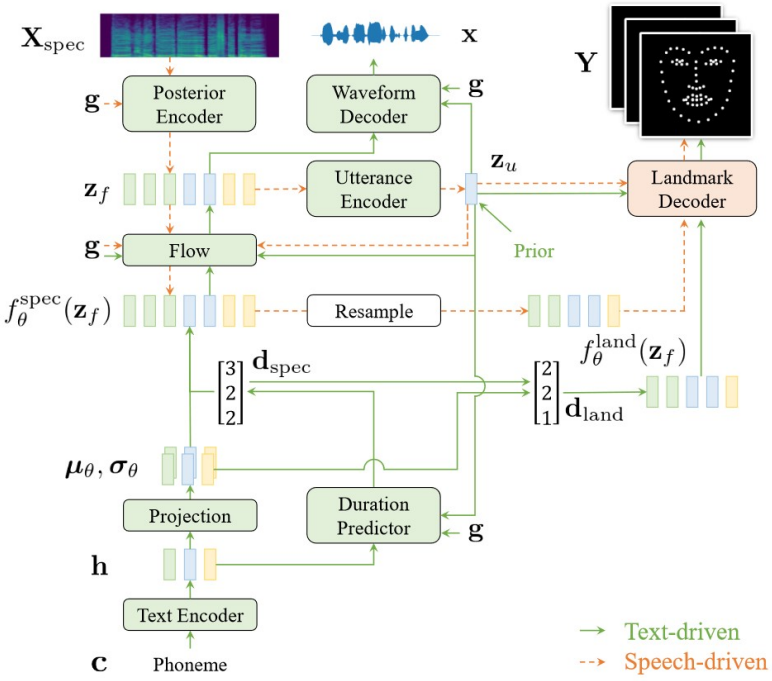


Figure 4.6: UniFLG consists of two components: a TTS based on the end-to-end VITS framework (Kim et al. 2021a), and a landmark decoder, which generates facial landmarks from the common representation of text and speech. UniFLG simultaneously generates speech and facial landmarks during text-driven inference, and it generates facial landmarks without using textual information during speech-driven inference. Figure from Mitsui et al. (2023).

4.4 Evaluation Approaches

For image generators, the quality of the generated output is usually assessed with the inception score, a metric that matches human perception and captures both diversity and realism. However, the *Fréchet Inception Distance* (FID), introduced by Heusel et al. (2017), has become much more prevalent. In time series, such as sound, *Dynamic Time Warping* (DTW) is a metric that can be used to measure similarity (Müller 2007). In the context of TTS, we can use DTW to judge the distance between the ground truth and the generated waveform. DTW is ideal for speech processing because it isn't affected by misalignment or variability in speed.

The sources of variation in talking head videos are highly coupled and can either be intrinsic (subject-related) or extrinsic (due to a moving camera or background). Thus evaluation is a challenging task, which is why most published research that involves speech synthesis resorts to subjective qualitative metrics to evaluate synthesized results, like *Mean Opinion Score* (MOS) or grader preference. These evaluations are often tedious and unreproducible. Quantitative metrics are usually error estimations using the ground-truth signal, like *Mean Square Error* (MSE) and other statistical measures. However, they may not be able to clearly specify whether the scores are meaningful and may not correspond to how humans perceive and judge generated video frames. For example, a slight misalignment between two frames may be unnoticeable to a human, but results in very large MSE between them. It is clear that perceptual metrics are necessary in order to benchmark such models. TTS also suffers from the one-to-many mapping problem, since the same text can be uttered in many different ways. Speaker identity, prosody and emotion are just some of the few sources of variation that can completely change the speech waveform. Consequently, waveform generation is uncertain and its evaluation is even trickier. Talking head videos need to be evaluated on four important aspects:

- Identity preservation of the reference person

- Visual quality
- Lip synchronization with the reference audio
- Natural and spontaneous motion of the head

Chen et al. (2020a) perform an extensive survey on talking head evaluation and provide several quantitative metrics for each of the aforementioned aspects. For instance, the distance of identity vectors extracted between frames is an index of identity preservation. Regarding visual quality, metrics like SSIM (Wang et al. 2004), SNR, and inception score are examined. A contrastive loss between audio and visual features can evaluate the lip synchronization. Finally, similar distances that measure perceptual quantities like emotional expression and blinking are introduced.

Chapter 5

Proposed Method: Neural Text to Articulate Talk

Contents

5.1 Motivation

The goal of this Diploma Thesis is to propose a novel deep learning model that creates talking face videos based on a text transcription. The model was conceived as a system for creating animated avatars to be used in virtual assistants, educational videos, accessibility tools, and human-machine interfaces. Thus, its development, architecture, and training data are all targeted towards producing videos with neutral facial expression and speech intonation. Such a model would be ideal for extending existing conversational agents that have recently exploded in popularity, due to the success of large language models (Zhang et al. 2023). Our system can provide an animated avatar that utters the conversational agent's response, paving the way towards a more natural mode of human-machine interaction (Lan et al. 2023). It can also contribute to building simpler animation pipelines for digital media professionals.

Since we aim to develop a neural network model that can generate a lifelike conversational avatar uttering natural language text, our main focus is on the realism of lip articulation, as well as the synchronization and naturalness of audio and visual streams. We decided to name our model "Neural Text to Articulate Talk", which we abbreviate to *NEUTART*.

As was established in Chapter 3, modeling the human face as a three-dimensional object instead of an two-dimensional image can better capture its variability, which is crucial for the perception of realism. Hence, a generative model of human faces that uses a three-dimensional representation can achieve better expressiveness in the generated output than a similar model operating directly on images. The mapping from the rendered 3D representation to a photorealistic image can be viewed as an image-to-image translation task, which is a well-studied problem (Isola et al. 2017).

The synchronization of facial movements and the uttered speech is equally essential for the talking head's realism. Of course, the natural language text, while being a sequence of characters, is not really a temporal signal. As a result, synchronization does not need to be enforced between the input and output of the system, only between the audio and visual modalities of the output. As we described in the previous Chapter, many models employ a pipelined approach of a TTS generator and an audio-driven video generator, which may introduce a temporal lag between the speech waveform and the talking video. This is why works like SyncNet (Chung and Zisserman 2017) focus on lip-syncing audio and visual streams of people speaking. Similarly, talking head generators like Prajwal et al. (2020), Cheng et al. (2022), Park et al. (2022), and Guan et al. (2023) formulate the video generation process as a lip-sync task. We can completely avoid this pitfall by generating the video and speech in parallel, instead of sequentially.

From the above analysis, we can extract two major requirements that shape our system's architecture. First of all, we would like our model to internally use three-dimensional modeling of the human face, which can be leveraged as an intermediate representation for creating photorealistic videos. Secondly, we would like to generate the audio and visual elements in a parallel way, without enforcing synchronization at the output.

These two requirements naturally lead us to a simple, but powerful design concept. The simultaneous generation of speech and a dynamic 3D talking head, with the latter being used to drive a conditional rendering model that synthesizes colored video frames. This architecture separates the task of text-driven photorealistic talking face generation into two subtasks: audiovisual 3D talking face generation and 3D-based photorealistic generation. We can leverage the speech-informed 3D reconstructions by SPECTRE (Filntisis et al. 2023) in order to train the audiovisual component with highly accurate talking head reconstructions, which can allow us to achieve unparalleled realism of lip movements. An abstract, high-level depiction of the architecture is depicted in Figure 5.1. It consists of an *audiovisual module* and a *photorealistic module*, which are thoroughly analyzed in the following Subsections.

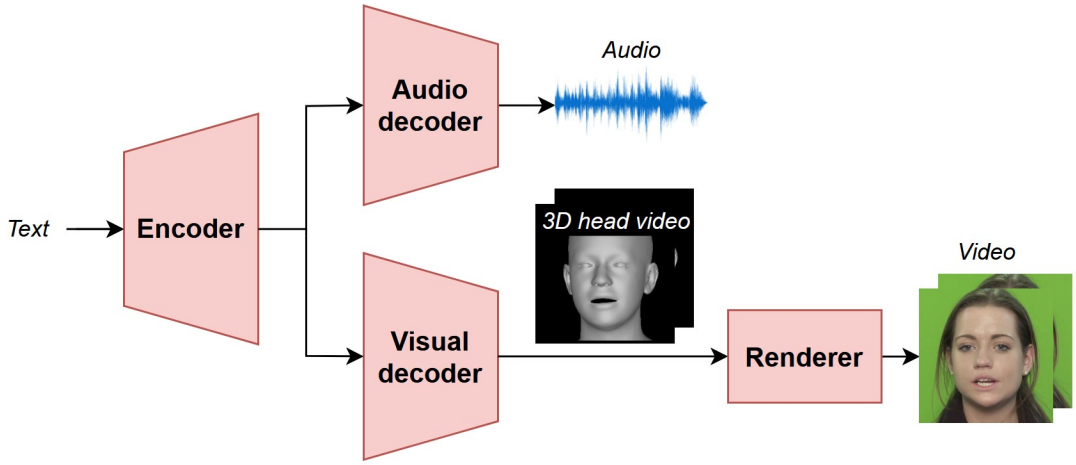


Figure 5.1: Our model uses text features to simultaneously generate the audio and visual streams of a talking face. The audio stream generates the speech waveform. The visual stream predicts the 3D reconstruction of the talking face, which is used as a condition to render the photo-realistic video. Since the audio and visual generation streams are driven by the same input, synchronization and coherence between them is implicitly enforced.

5.2 Data Preprocessing

Audiovisual speech synthesis datasets consist of talking head video clips accompanied by text transcriptions. In order to train each module, we need to preprocess the data in order to extract the information that each module requires. For instance, we need to perform text-to-audio alignment between the text transcription and the audio of the talking head clip. We also need to isolate the subject's face and perform 3D reconstruction of their head. We explain the different datasets used for our experiments in the next Chapter. For now, we present an overview of the preprocessing steps.

Face Detection

We begin by processing the visual stream. First of all, we perform face detection using a publicly available implementation of MTCNN (Zhang et al. 2016). This work uses a cascaded architecture with three stages of convolutional networks to predict the faces in a coarse-to-fine manner, by quickly rejecting the background regions in the fast low resolution stages, while carefully evaluating a small number of candidates in the last high resolution stage. We use this method to obtain a square bounding box of the face region in each frame of the video. We slightly enlarge the detected bounding box to account for the whole head, crop the image around it, and resize the interior to 256×256 pixels. For in-the-wild videos, face detection may return false positives in the background, which we can avoid by simply selecting the largest bounding box.

Face Segmentation

To make our method background-agnostic, we constrain the renderer to predict only the face interior. To this end, we employ a face segmentation network that predicts a mask, which we use to remove the non-face areas of the image. We employ the FSGAN (Nirkin et al. 2019), a simple convolutional architecture based on the U-Net image segmenter (Ronneberger et al. 2015). We obtain a binary mask of the face and we softly erode it to smooth its boundary. Thus, we can only keep the face interior by element-wise multiplying the given cropped image with this face mask.

Head Reconstruction

We use the SPECTRE method introduced by Filntisis et al. (2023), and presented more thoroughly in Subsection 3.3.2, to encode the face images in the FLAME 3DMM (Li et al. 2017), as well as estimate the scene parameters. We then render the 3D head mesh into an image, which is used to condition the renderer. The input to the conditional neural renderer consists of the concatenation of the rendered 3D face geometry image, as well as the *Normalized Mean Face Coordinate* (NMFC) rendering, introduced by Doukas et al. (2021b). The NMFC is a semantic representation of the head. It depends not only on the 3D head reconstruction and camera parameters of the current frame, but also on the 3DMM template. The 3D spatial coordinates of the normalized template are used as constant color values, adding texture to the rendered head which is reconstructed from the current frame.

Landmark Detection

We use the popular FAN detector (Bulat and Tzimiropoulos 2017), to obtain 68 facial landmarks for each frame. The need for detecting these 2D landmarks is threefold.

1. We use the eye-specific landmarks to locate the two eye pupils.
2. We perform 2D face alignment based on a template set of 68 landmarks that are centered in a square frame.
3. We use the mouth-specific landmarks to crop the subject's face around the mouth.

Following the implementation of NED (Paraperas Papantoniou et al. 2022), we estimate the eye centers as additional landmarks, by taking the weighted center of mass of pixel coordinates enclosed in each polygon of eye landmarks. The weights are simply the inverse of pixel intensities, thus biasing the center towards the actual eye pupil, which is expected to be darker than the rest of the eye (Saragih et al. 2011). We then create eye images that provide the face renderer with information about the eye gaze. These consist of merely two red circles that are drawn on the pupil locations.

Face Alignment

We perform face alignment order to boost the renderer's generalization ability, by bringing all faces to a common reference. This is similar to correspondence establishment in 3DMMs, which was shown in Figure 3.2. Aligning all faces facilitates the learning process of this network, by removing any bias that might occur from the different poses or positions. To this end, we use the Umeyama least-squares estimation (Umeyama 1991) of the optimal 2D similarity transformation matrix between the 68 extracted landmarks and the corresponding landmarks of a mean face template, per video frame. The masked face images, as well as the NMFC, shape, and eye-gaze images are then warped according to this transformation, since they are used as input to the renderer during training.

We present the video processing pipeline with one frame from our experiments in Figure 5.2.

Text-to-Audio Alignment

Having processed the visual stream, we move on to the audio stream. As we saw in Subsection 2.1.2, the audio waveform has to be aligned with the text's phoneme sequence. We use the Montreal Forced Aligner (McAuliffe et al. 2017), which converts a plain text transcription to the corresponding sequence of phonemes paired with their start and end timestamps in the audio. Thus, we can also extract the duration of each phoneme, which is crucial for accurate duration prediction, since the phoneme sequence needs to be appropriately expanded before it is mapped to a spectrogram.

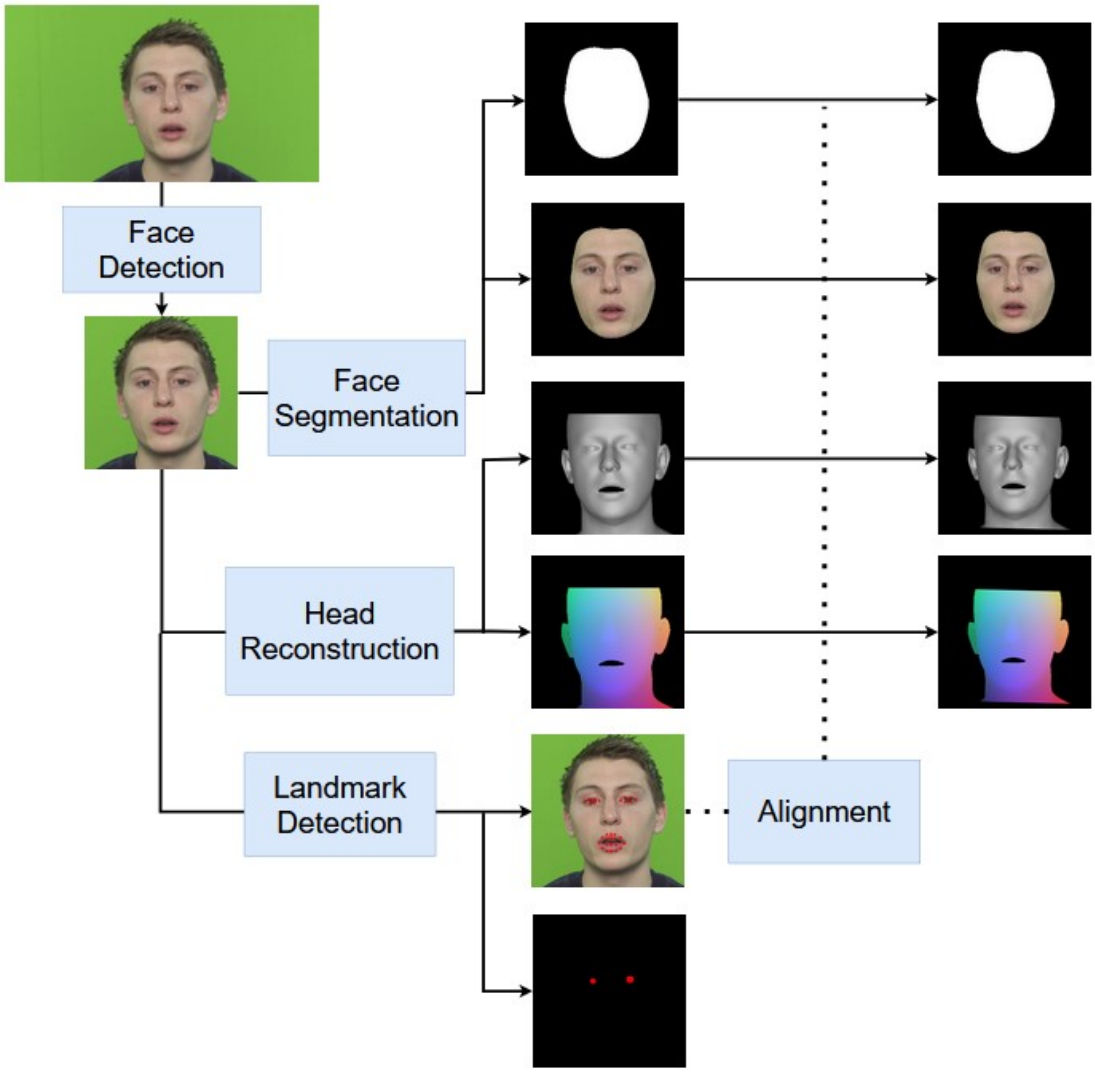


Figure 5.2: Visualization of the video preprocessing pipeline for one frame. The face is detected and the original frame is cropped around it. Then, a mask image is created to isolate the face interior. The head is reconstructed in 3D, while 2D landmarks are predicted from the face image, spanning borders of the eyes and mouth, visualized at the bottom of the figure. Finally, a transformation is calculated in order to bring the landmarks in correspondence with a landmark template. This transformation is applied to all extracted images.

Audio Parameters Calculation

We use the established WORLD framework (Morise et al. 2016) to calculate the fundamental frequency of the speech signal per audio frame. Furthermore, the spectrogram is calculated by projecting the STFT of the speech signal onto mel frequency bins. We also calculate the $L2$ norm of each STFT frame as the signal's energy. For more information about the mel spectrogram, please revisit Subsection 2.1.3.

5.3 Audiovisual Module

5.3.1 Architecture

The task of speech generation has been tackled quite successfully by TTS systems, as we saw in Chapter 2. To build the audiovisual module, we choose to leverage a robust TTS architecture and extend it to incorporate a 3DMM generation component. One prominent TTS system that is frequently used in recent publications such as the ones by Łańcucki (2021) and Min et al. (2021) is FastSpeech 2 (Ren et al. 2020). This system uses transformers (Vaswani et al. 2017) to convert a sequence of phonemes into a sequence of mel characteristics, from which any pre-trained vocoder can produce the speech waveform. The architecture of FastSpeech 2 is depicted in Figure 2.3, showing the encoder-decoder transformer scheme. We extend this spectrogram prediction architecture by also predicting a vector of 3DMM coefficients per audio frame. Thus, not only are the vector sequences of audio and visual characteristics synchronized by design, but also TTS is converted to a bimodal task.

To be precise, our own system's architecture follows the encoder-decoder scheme of FastSpeech 2, but adds a second decoder for the generation of FLAME 3DMM coefficients. Both the *audio decoder* and the *visual decoder* are jointly driven by the output of the text encoder, which is advantageous in the following ways:

1. The correlation and highly complex interplay between audio and visual streams that occurs during speaking is built into the network.
2. We construct a direct mapping of linguistic features into uttered speech, which essentially models the process of a human reading out loud a piece of written text. Thus, the model's learned feature space is a neural representation of audiovisual speech.
3. We model the one-to-many mapping from audio to video during speech, by implicitly considering the joint distribution $p(\texttt{audio}, \texttt{video}|\texttt{text})$ instead of the rougher approximation by chaining $p(\texttt{audio}|\texttt{text})$ and $p(\texttt{video}|\texttt{audio})$.
4. We avoid the redundancy of 2-stage architectures that drive the talking face generator with features from an audio feature extractor. This feature extractor encodes the TTS system's synthesized speech into another speech-related representation, whereas the TTS system already used intermediate speech features extracted from the text. Encoding the speech signal twice is computationally redundant and does not offer any advantage in terms of generation quality.
5. We also avoid the potential dataset mismatch and unavoidable error accumulation of cascaded approaches, especially those that do not retrain the TTS component. The utilization of out-of-the-box TTS audios for driving a talking face generator is likely to introduce artifacts, since the latter has to be trained with real audios from video samples. Regardless of whether the training of the two stages is performed with the same data, TTS artifacts are amplified in the speech-to-talking-head inference. The two speech feature spaces learned by each stage may have completely different dimensionality and topology, so the learned projection between them is bound to introduce errors.

Since we use SPECTRE for the 3D reconstructions, the faces are modeled with the FLAME 3DMM (Li et al. 2017), which decouples a 3D head mesh of 5023 vertices into vectors of shape $\vec{\beta}$, expression $\vec{\psi}$ and pose $\vec{\theta}$ parameters. We have introduced the FLAME formalism in Section 3.2.2.

The pose vector includes the 3 jaw articulation parameters $\vec{\theta}_{jaw}$, modeling the jaw as a spherical joint with 3 degrees of freedom in space (Craig 2021). Using the FLAME representation, we model the facial expression and movements during speech using the 3 jaw pose and 50 expression parameters, concatenated to a vector $\vec{x} \in \mathbb{R}^{53}$.

Formally, the audiovisual module is a neural network $\mathcal{A}$ the learns the transformation from an input sequence $p_{1:L}$ to the output sequences $\vec{x}_{1:N}$ and $\mathbf{F} = \vec{y}_{1:N}$.

$$\mathcal{A}(p_{1:L}) \rightarrow (\mathbf{F}, \vec{x}_{1:N}) \tag{5.1}$$

The sequence $p_{1:L} = (p_1, ..., p_L),\ p_i \in P$ contains L lexically stressed items from the ARPA-bet phoneme set P described in Table 2.1. We also need to use an additional phoneme that corresponds to silence. Regarding the outputs, $\vec{y}_{1:N} = (\vec{y}_1, ..., \vec{y}_N),\ \vec{y}_n \in \mathbb{R}^{80}$ is the sequence of mel characteristic vectors that makes up the spectrogram $\mathbf{F}$ when concatenated. Similarly, $\vec{x}_{1:N} = (\vec{x}_1, ..., \vec{x}_N),\ \vec{x}_n \in \mathbb{R}^{53}$ is the sequence of 3DMM coefficient vectors. The length N of the predicted sequences is longer than the length L of the input, since each phoneme can span more than one audio frame.

The outputs of the audiovisual module are converted to audio and visual elements through pretrained models. The vocoder $\mathcal{V}$ creates the continuous speech signal $s(t) \in \mathbb{R}$.

$$\mathcal{V}(\mathbf{F}) \rightarrow s(t) \tag{5.2}$$

Obviously, the signal is digitally sampled, but we find that denoting it as continuous is more intuitive in this context, since the audio samples are vastly larger in number than the video frames.

Similarly, FLAME converts each 3DMM vector $\vec{x}_n$ into a 3D mesh which is used as input to the photorealistic module. Since we only model speech-related variation, we use the prediction $\vec{x}_n$ along with the ground truth parameters for shape and head pose as input to the FLAME model, whose formulation was described in Equation 3.7.

Overall, the audiovisual module needs to predict 80 mel channels and 53 3DMM channels per audio frame. The module consists of the following submodules:

- **Phonemizer:** This first submodule takes a plain English text as input and converts it to phoneme sequences according to the CMU pronunciation dictionary. For OOV words, it uses a simple prediction network to infer their phonetic spelling (Park and Kim 2019). Note that this component is used only for inference, since the plain text needs to be converted to phonemes. For training, this has been already done during preprocessing.
- **Encoder:** The encoder projects the sequence of phoneme indices into a sequence of phoneme embeddings, which is processed by a 4-layer transformer.
- **Variance Adaptor:** This submodule is used to add variance to the text encodings. The information modeled consists of three audio features, namely pitch (fundamental frequency), spectrogram energy, and phoneme duration. The variance adaptor expands each phoneme encoding according to its duration in mel frames, then adds pitch and energy embeddings to the phoneme encodings.
- **Audio Decoder:** This submodule consists of a 6-layer transformer and a linear layer, which project the intermediate features into the spectrogram.
- **Vocoder:** We use the pretrained HiFi-GAN universal generator (Kong et al. 2020a) in order to convert the mel spectrogram into a speech timeseries.
- **Visual Decoder:** Similar to the audio decoder, a 4-layer transformer and a linear layer to convert the intermediate features into timeseries of pose and expression coefficients. The audio and visual decoders were chosen to have different depths due to the different dimensionality of the data they need to model. The former needs to predict a value for 80 mel bands, while the latter needs to predict 53 independent 3DMM components.

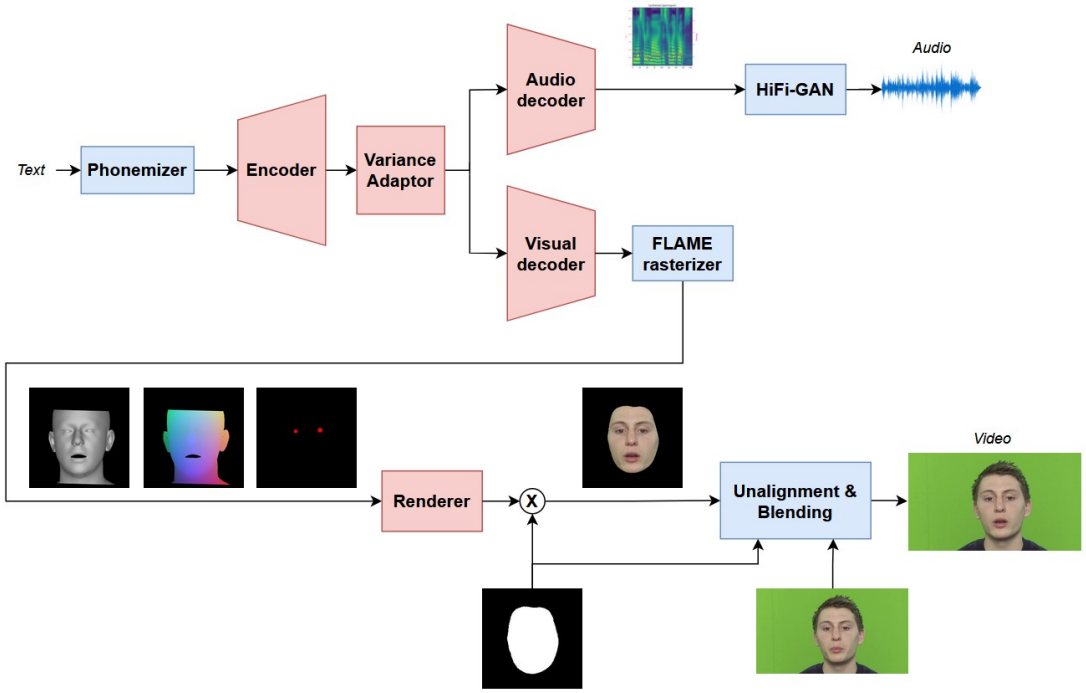


Figure 5.3: An overview of the inference process of NEUTART, starting from a plain text transcription, and producing a photorealistic talking head video. The audiovisual module (top half) converts the text to a phoneme sequence, which is jointly mapped to the spectrogram and 3D head parameters, which are respectively converted to audio and 3D head by the HiFi-GAN vocoder and FLAME rasterizer. Then, the 3D head video is used by the photorealistic module (bottom half) to generate the subject's face, which is swapped with the face from a reference video (for simplicity, we only show one video frame, despite depicting the entire spectrogram and audio sequence above). The networks that our method needs to optimize are shown in pink, while frozen components and deterministic processing stages are shown in light blue. Note that only the background from the reference video is used.

- **FLAME rasterizer:** A FLAME decoder projects the vector x_n of pose and expression coefficients to a 3D mesh, per video frame of index n. From each 3D mesh, a shape as well a NMFC image are rasterized, which are used to drive the renderer.

The above components and the information flow among them are depicted in the top half of Figure 5.3.

5.3.2 Training

Each training sample consists of the following information from the dataset, per utterance:

- The **phoneme** sequence $p_{1:L}$.
- The average **pitch** per phoneme $\hat{f}_{1:L}$.
- The average **energy** per phoneme $\hat{e}_{1:L}$.
- The **duration** of each phoneme $\hat{d}_{1:L}$.
- The ground truth mel **spectrogram** $\hat{\mathbf{F}}$.
- The **3DMM** coefficient vectors of the SPECTRE reconstruction $\hat{\vec{x}}_n$.
- A crop of the original video around the **mouth** $\hat{\mathbf{I}}^M_{1:N}$.

- The ground truth **camera** parameters per video frame $\vec{c}_{1:N}$.

The overall loss function $\mathcal{L}_{av}$ is defined as the sum of the following loss terms, per utterance:

- **Pitch, Energy and Duration losses:** Following FastSpeech 2, these losses aim to adjust the predictions of the variance adaptor. They are defined as the MSE between the ground truth value and the predicted one. For instance, we formulate the pitch loss:

$$\mathcal{L}_{pitch} = \mathbb{E}_l[||\hat{f}_l - f_l||_2^2] \tag{5.3}$$

  where f_l is the predicted pitch value for the phoneme at index l.

- **Spectrogram loss:** Again, we follow FastSpeech 2 and optimize the audio decoder's output, which is the spectrogram, using their $L1$ distance. Formally:

$$\mathcal{L}_{mel} = ||\hat{\mathbf{F}} - \mathbf{F}||_1 \tag{5.4}$$

  where $\mathbf{F}$ is predicted spectrogram. Equation 5.4 slightly abuses the notation of $L1$ norm, since it refers to images. In practice, the distance is computed on flattened versions of the spectrograms, so that they are vector-shaped.

- **3DMM loss:** A simple MSE between the predicted and the ground truth 3DMM parameters.

- **Gradient loss:** We penalize the temporal gradient of the 3DMM coefficient vectors, in order to enforce smoothness on the prediction, with the loss term:

$$\mathcal{L}_{grad} = \mathbb{E}_n[||\vec{x}_{n+1} - \vec{x}_n||_2^2] \tag{5.5}$$

- **Flow loss:** We also use the temporal differences loss that Hussen Abdelaziz et al. (2020) use in their multimodal taking face generator. We prefer to call it *flow loss*, to avoid confusion with the aforementioned gradient loss, and formulate it as:

$$\mathcal{L}_{flow} = \mathbb{E}_n[||(\hat{\vec{x}}_{n+1} - \hat{\vec{x}}_n) - (\vec{x}_{n+1} - \vec{x}_n)||_2^2] \tag{5.6}$$

  where $\hat{\vec{x}}_n$ is the actual target value for the FLAME vector and $\vec{x}_n$ is the predicted vector at frame n, for a total of N frames.

- **Lipreading loss:** Following the speech-informed training of SPECTRE, we use the intermediate features from a pretrained lipreading model (Ma et al. 2022) in order to capture the patterns of lip movements while speaking. The ground truth video and the generated 3D mesh video are cropped around the mouth, then feature vectors are extracted for each of their frames. We use their cosine distance as:

$$\mathcal{L}_{lip} = \mathbb{E}_n \left[1 - \frac{\hat{\vec{f}}_n \cdot \vec{f}_n}{||\hat{\vec{f}}_n||\ ||\vec{f}_n||} \right] \tag{5.7}$$

  where $\hat{\vec{f}}_n,\ \vec{f}_n \in \mathbb{R}^{512}$ are the feature vectors at frame n for the ground truth and predicted video, respectively. The inputs to the lipreading network are visualized in 5.4.

- **Expression regularization loss:** As Filntisis et al. (2023) note, the lipreading loss needs an additional constraint on the magnitude of the expression coefficients, otherwise they are prone to oscillations. Thus, along with $\mathcal{L}_{lip}$, we use the following $L2$ regularization loss:

$$\mathcal{L}_{reg} = 10^{-3}\mathbb{E}_n[w_n||\vec{\psi}_n||_2^2] \tag{5.8}$$

  with w_n empirically found by SPECTRE's authors to yield better results for:

$$w_n = \begin{cases} 1, & ||\vec{\psi}_n||_2^2 < 40 \\ 2, & ||\vec{\psi}_n||_2^2 > 40 \end{cases} \tag{5.9}$$

Figure 5.4: Example frames from the video of an utterance cropped around the mouth, which are used as input to the pretrained lipreader. The first row depicts real frames from the training video, while the bottom row shows the predicted 3D textured head's mouth for that utterance, during training. The lipreading loss aims to make these mouth positions as similar as possible, contributing positively to the realism of the synthetic output. We chose to present four random frames, instead of consecutive, so that we visualize a wider variety of visemes. The videos are cropped to 88×88 resolution, in order to match the training of the lipreading model.

Overall, the loss function that we use to optimize the audiovisual module can be written as follows.

$$\mathcal{L}_{av} = \mathcal{L}_{pitch} + \mathcal{L}_{energy} + \mathcal{L}_{dur} + \mathcal{L}_{mel} + \mathcal{L}_{3DMM} + \mathcal{L}_{grad} + \mathcal{L}_{flow} + \mathcal{L}_{lip} + \mathcal{L}_{reg} \tag{5.10}$$

We have incorporated any scaling weights into each constituent loss, for cleaner notation, thus the losses are added without any weights in the above expression. The overall training procedure of the audiovisual module, highlighting the network inputs, outputs, and losses, is presented in Figure 5.5.

One implementation detail is that the 3DMM and mel sequences have different temporal resolutions. The 3DMM vectors are extracted per frame, at 25 frames per second (fps), while the latter are extracted per mel frame. The audios are sampled at 22.05 kHz, and we use a temporal window of 1024 samples, with a hop length of 256 samples. This corresponds to $\frac{22050}{256} \approx 86.13$ mel frames per second. In order to preserve audio quality, we upsample the 3DMM sequences to match the audio fps, so that the two decoders can jointly learn to generate the audio and visual streams. After the visual decoding, the predicted 3DMM vector sequences are downsampled back to 25 fps, since loss calculation requires them to match the ground truth lengths. Both resampling operations are performed using linear interpolation.

5.4 Photorealistic Module

5.4.1 Architecture

This module's main component is a GAN renderer $\mathcal{R}$ which is trained to predict face crops using the rendered 3D meshes as input. We follow the face renderer architecture that was proposed in Head2Head++ (Doukas et al. 2021b) and adapted in Neural Emotion Director (Paraperas Papantoniou et al. 2022). This approach allows us to modify the visual content of an input video featuring a speaking person. The neural renderer is implemented with a convolutional architecture that tackles an image-to-image translation task based on GANs.

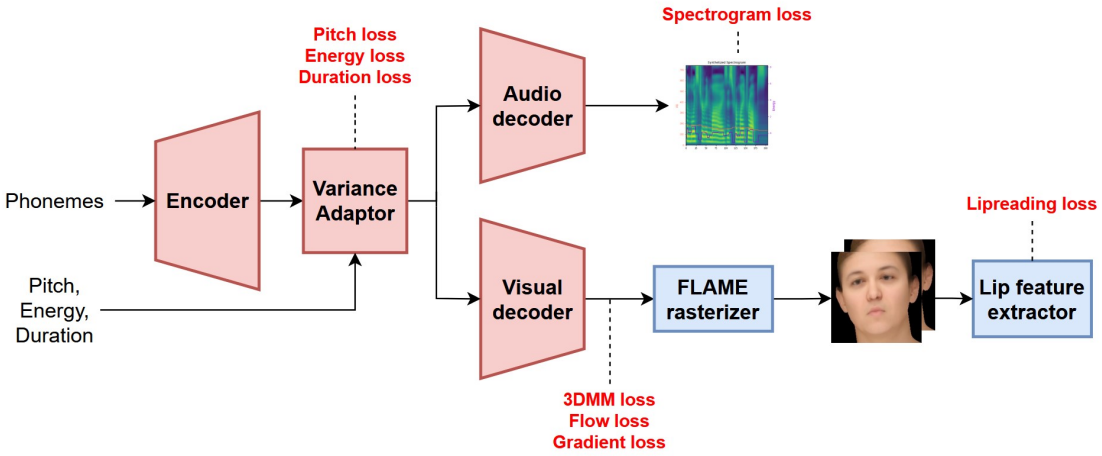


Figure 5.5: The training procedure of the audiovisual module requires the phoneme sequence corresponding to an utterance's transcription, as well as the average pitch, energy, and duration for each phoneme in the sequence. The variance adaptor is trained to predict the pitch, energy, and duration per phoneme, using the ground truth values for training and predicted values for inference. The output of the audio decoder is the mel spectrogram, which is used to calculate the spectrogram loss. Similarly, the output of the visual decoder is the 3DMM sequence, which is mapped to a 3D mesh using FLAME, then rasterized in an image, from which lipreading features can be computed. The ground truth values required by some of the losses are implied inputs and not shown in this diagram. Trainable components are shown in pink color, frozen ones in light blue. Notice that there is both audio and visual supervision, treating visual speech synthesis as a bimodal task.

Formally, the renderer predicts a video frame $\mathbf{I}_n \in \mathbb{R}^{W \times H \times 3}$, based on the image $\mathbf{S}_n \in \mathbb{R}^{W \times H \times 3}$ of the 3D shape, the NMFC image $\mathbf{N}_n \in \mathbb{R}^{W \times H \times 3}$, the eye image $\mathbf{E}_n \in \mathbb{R}^{W \times H \times 3}$, as well as the previous two frames. For simplicity, we denote the channel-wise concatenation of $(\mathbf{S}_n, \mathbf{N}_n, \mathbf{E}_n)$ as $\mathbf{X}_n \in \mathbb{R}^{W \times H \times 9}$. As we described in Section 5.2, both the width W and the height H are 256. Thus, the renderer learns the following mapping, per video frame:

$$\mathcal{R}(\mathbf{X}_n, \mathbf{I}_{n-1}, \mathbf{I}_{n-2}) \rightarrow \mathbf{I}_n \tag{5.11}$$

For the start of the generation, we simply use $\mathbf{I}_{-2} = \mathbf{I}_{-1} = \mathbf{I}_0$. Concatenating the sequence of generated frames along the time dimension yields the output video $\mathbf{I}_{1:N}$, which is blended with some reference video of the subject in order to fill the background.

The GAN training setup consists of the aforementioned convolutional generator $\mathcal{R}$, as well as an image discriminator $\mathcal{D}$ and a dedicated mouth discriminator $\mathcal{D}_M$, aiming to enhance realism in the mouth area. Since we are only interested in synthesizing the face region, the generator's output is masked with the extracted face mask $\mathbf{M}_n \in \{0,1\}^{W \times H}$. Following the implementation of Paraperas Papantoniou et al. (2022), the generator is built similarly to Vid2Vid (Wang et al. 2018a), while the discriminators adopt their architectures from Pix2PixHD (Wang et al. 2018b).

5.4.2 Training

The renderer is trained to reconstruct the masked face from the original RGB frame, conditioned on the shape and NMFC images. It employs a GAN-based adversarial loss (Goodfellow et al. 2014), as well as a specialized mouth discriminator for improved realism in the lip area.

As already mentioned, we present the datasets used for our experiments in the next Chapter, but we now present an overview of the preprocessed training data. The photorealistic module is also trained per utterance video clip. Thus, each training data instance consists of the following items:

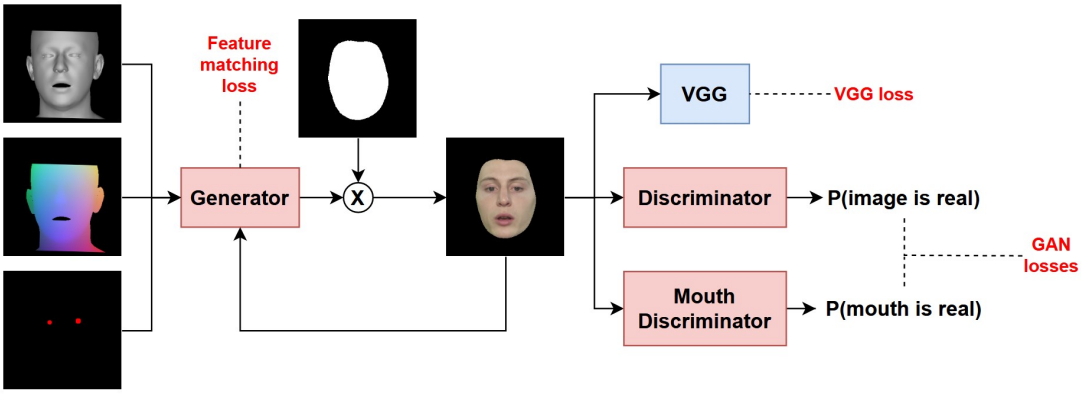


Figure 5.6: A snapshot of the training procedure of the photorealistic module for one video frame. The training follows the adversarial protocol employed in GANs, using a conditional generator (renderer $\mathcal{R}$) and two discriminators. The conditional inputs to the generator are the aligned shape, NMFC, and eye images, as well as the two previous faces, depicted with a simple feedback loop. Its output is masked using the ground truth face mask, and serves as the input to the two discriminators. Mouth cropping before the mouth discriminator is implied, and so are the ground truth faces which are used to calculate the losses.

- The **shape** video $\mathbf{S}_{1:N}$.
- The **NMFC** video $\mathbf{N}_{1:N}$.
- The **eye** gaze video $\mathbf{E}_{1:N}$.
- The face **mask** video $\mathbf{M}_{1:N}$.
- The **face** interior video $\hat{\mathbf{I}}_{1:N}$.
- The **mouth** video $\hat{\mathbf{I}}^{M}_{1:N}$.

We have presented an overview of GAN training in Section 1.2. The training objectives $\mathcal{L}_{ph}$ for the photorealistic module, following NED, Vid2Vid and Pix2PixHD, are as follows.

- **Adversarial loss:** For the adversarial training, we employ the Least Squares GAN loss proposed by Mao et al. (2017), which improves training stability. The game objective for the generator $\mathcal{R}$ is to make the discriminator predict a high probability that a generated sample is real:

$$\mathcal{L}^{\mathcal{R}}_{GAN} = \frac{1}{2}\mathbb{E}_n[(\mathcal{D}(\mathbf{X}_n, \mathbf{I}_n) - 1)^2 + (\mathcal{D}_M(\mathbf{X}^M_n, \mathbf{I}^M_n) - 1)^2] \tag{5.12}$$

  where $\mathbf{I}_n$ is the current output of the generator as described in Equation 5.11, and $\mathbf{I}^M_n$ is the aforementioned image cropped around the mouth. In contrast, each discriminator needs to predict low probability for fake data, and high for real data. For instance, the discriminator seeing the entire synthesized face needs to optimize:

$$\mathcal{L}^{\mathcal{D}}_{GAN} = \frac{1}{2}\mathbb{E}_n[(D(\mathbf{X}_n, \hat{\mathbf{I}}_n) - 1)^2 + \mathcal{D}(\mathbf{X}_n, \mathbf{I}_n)^2] \tag{5.13}$$

- **VGG loss:** This perceptual loss uses the *Visual Geometry Group* (VGG) network, introduced by Simonyan and Zisserman (2015), to extract visual features and find the distance between them. It can be written as:

$$\mathcal{L}_{VGG} = \mathbb{E}_n \left[\sum_i \frac{1}{M_i} ||\mathcal{F}_i(\hat{\mathbf{I}}_n) - \mathcal{F}_i(\mathbf{I}_n)||_1 \right] \tag{5.14}$$

  where $\mathcal{F}_i$ is the i-th layer of the VGG network, with a total of M_i elements.

- **Feature matching loss:** This is another perceptual loss that penalizes the differences between feature maps for real and generated images. It can be expressed similarly to Equation 5.14, but operating on the discriminator layers $\mathcal{D}_i$:

$$\mathcal{L}_{FM} = \mathbb{E}_n \left[\sum_i \frac{1}{M_i} ||\mathcal{D}_i(\hat{\mathbf{I}}_n) - \mathcal{D}_i(\mathbf{I}_n)||_1 \right] \tag{5.15}$$

Overall, the loss function that we use to train the generator $\mathcal{R}$ is:

$$\mathcal{L}_{ph} = \mathcal{L}^{\mathcal{R}}_{GAN} + 10\mathcal{L}_{VGG} + 10\mathcal{L}_{FM} \tag{5.16}$$

The weighting parameters were selected following the aforementioned previous works. Notice that the discriminators are only involved in the adversarial loss, thus their training objective is simply Equation 5.13. The overall training procedure of the photorealistic module is depicted in Figure 5.6.

It is important to note that the renderer is unimodally trained as an image generator, similarly to how the vocoder is pretrained with only audio supervision. Also, in contrast to NED that used the DECA method (Feng et al. 2021) for 3D reconstruction, we have employed SPECTRE, which focuses on visual speech-preserving 3D reconstruction. The same 3D reconstructions are used as ground truth values for training the visual decoder in our audiovisual module, ensuring consistency between them. The above components and the information flow among them are depicted in the bottom half of Figure 5.3. The two modules are coupled during inference, but are separately trained on the same data, due to the heavy computational requirements of the neural renderer used in the photorealistic module.

Chapter 6

Experiments

Contents

Dataset	Hours	Subjects	Sentences	Environment
TCD-TIMIT	11.1	62	6.9k	Lab
LIPS2008	0.25	1	280	Lab
HDTF	15.8	300	10k	In-the-wild

Table 6.1: Information about the datasets that we experimented on. The HDTF dataset is richer, but at the same time more challenging due to the in-the-wild recording conditions. It is also not tailored for speech synthesis, unlike the TCD-TIMIT dataset which is designed to be rich in phoneme and viseme coverage. The LIPS2008 dataset is very low-resource, as it was released for a visual speech synthesis challenge.

6.1 Datasets

6.1.1 Lab Conditions

TCD-TIMIT

Most large-scale audiovisual datasets are targeted towards audiovisual speech recognition (Chung et al. 2017), or speech-driven talking face generation, often targeting emotional expression (Cao et al. 2014; Wang et al. 2020). Thus, they usually focus on the visual quality of the recordings, their abundance, as well as the subject variability. However, for training a text-driven audiovisual model, we need a dataset with high-quality audio as well, preferably recorded in lab conditions. The most suitable such dataset that is publicly available is TCD-TIMIT (Harte and Gillen 2015). It consists of high quality video and audio recordings of 62 speakers, reading sentences from the established phonetically rich TIMIT corpus. The transcriptions are also available, alleviating the need for performing automatic transcription, which is bound to introduce errors in the training data.

LJSpeech

Nevertheless, the audio recordings are still relatively noisy compared to the recordings in pure TTS datasets. In order to mitigate this, as well as increase the robustness of speech generation, we use transfer learning (Yosinski et al. 2014) by initializing the audiovisual module's encoder and audio decoder with weights from a FastSpeech 2 model trained on the LJSpeech dataset (Ito and Johnson 2017). LJSpeech is an established dataset in speech synthesis, and consists of recordings of a single female speaker reading passages from non-fiction books, with a total duration of 24 hours. After that initialization, we train the audiovisual module using the audios and 3-dimensional face reconstructions of TCD-TIMIT clips. The photorealistic module is trained using the video frames of the same TCD-TIMIT clips.

LIPS

We also experimented with the low-resource LIPS2008 dataset, recorded in lab conditions, and presented in Theobald et al. (2008). This dataset was created to provide an evaluation benchmark for visual speech synthesis methods.

6.1.2 In-the-wild

HDTF

Furthermore, we conducted experiments using in-the-wild videos to demonstrate the capabilities of our method. We used clips from the High-Definition Talking Face (HDTF) dataset (Zhang et al. 2021), which consists of high quality online videos, mostly of American politicians giving speeches. Sample frames from each dataset are presented in Figure 6.1, while their technical characteristics are highlighted and compared their in Table 6.1.

Figure 6.1: Sample frames from the datasets that were used in this Thesis. The top row presents sample clips of different subjects of the TCD-TIMIT dataset, recorded in lab conditions. The middle row consists of frames depicting in-the-wild samples from HDTF, and is borrowed from Zhang et al. (2021). The bottom row depicts consecutive frames from an utterance of the LIPS2008 dataset.

YouTube videos

Finally, we experimented with completely unconstrained talking head videos that are publicly available on YouTube. We used those videos to only train the photorealistic module, and drove it with the audiovisual output from a TCD-TIMIT subject of the same gender, in order to examine its generalization capabilities.

6.2 Evaluation

We chose to evaluate our model only on samples from established and well-known datasets, in order to conduct fair and reproducible comparisons. We conducted qualitative and quantitative evaluations of our method and comparisons with recent state-of-the-art methods. As mentioned in the previous Section, we experimented with the large-scale, multi-speaker audiovisual dataset TCD-TIMIT (Harte and Gillen 2015), as well as some in-the-wild videos from the HDTF dataset (Zhang et al. 2021). The main experiments were conducted on TCD-TIMIT subjects, and those results are thoroughly analyzed in the next Subsections.

We first trained the audiovisual module in a multispeaker setting, following the multispeaker architecture of FastSpeech 2 (Ren et al. 2020), in order to benefit from the data abundance of the entire train split of the dataset. The total footage per speaker is less than 10 minutes, which is too low for training a speech synthesis model. Also, as already mentioned, we initialized the encoder and audio decoder with pretrained weights from the LJSpeech dataset. The visual decoder was randomly initialized, and all the models were trained for 50,000 iterations in each experiment. The trained multi-speaker model can be used as is, or fine-tuned to a particular identity for a few epochs, in order to create person-specific models. We observed that person-specific models yield more realistic results, so we fine-tuned the multispeaker audiovisual module on footage from single subjects, for 3,000 iterations. The photorealistic module is inherently person-specific, trained on RGB videos of each identity. We show some generated samples in Figure 6.2.

Figure 6.2: We present a few frames from various samples that we generated, with subjects from TCD-TIMIT and in-the-wild YouTube videos. Our method can produce photorealistic videos of talking heads, with lifelike lip articulation. On the left we include the word or phrase uttered in each set of frames. The top and bottom subjects are from TCD-TIMIT, while the middle ones are from the speech "Tell India's Story" and the channel "Reaction Therapy", respectively.

As discussed in Section 4.4, valuation for talking faces is a challenge, since statistical error measures do not correlate very well with human assessments (Chen et al. 2020a). We objectively evaluated our model using both statistical and perceptual metrics across different modalities, and also conducted a user study.

6.2.1 Compared Methods

We want to evaluate our model based on its audio and photorealistic video output, therefore we compare with recent methods that generate RGB video. As a result, we are not comparing our model against AVTacotron2 (Abdelaziz et al. 2021) or UniFLG (Mitsui et al. 2023), which are true audiovisual synthesizers, because their output is an untextured 3D rendering. Furthermore, both are proprietary implementations, while we would like to experiment with open-source works, for the sake of reproducibility.

Furthermore, the research conducted in this Diploma Thesis targets audiovisual speech generation from text. Since there are not any recent publicly available text-driven methods, the fairest option would be to compare against audio-driven models that output photorealistic video, but using synthetic audio to drive them. We choose to compare our model against the following popular lip-syncing models, with our method being abbreviated as NT in some Tables and Figures.

- Wav2Lip (Prajwal et al. 2020), abbreviated as W2L
- SadTalker (Zhang et al. 2022c), abbreviated as ST
- VideoReTalking (Cheng et al. 2022), abbreviated as VRT

These models were briefly described in Section 4.2, along with the analysis of the audio-driven lip-syncing mechanism. We sampled these models using audios from a FastSpeech 2 TTS model, with the same encoder and audio decoder architecture as our audiovisual module. This choice ensures the best possible fairness, since this TTS model and our audiovisual module have the same architecture and number of parameters, and are trained on the same data. Also, the compared methods are either one-shot or few-shot, meaning that they are not person-specific. Rather, they directly animate a given reference frame of the subject, or simply modify a given reference video by changing the subject's mouth. This offers them excellent generalization capabilities by design, however they are not fit for personalized highly-realistic avatars. We used real videos from TCD-TIMIT as reference during sampling from the aforementioned models.

6.2.2 Objective evaluation

The objective evaluation is performed on 3 randomly chosen subjects from the TCD-TIMIT dataset, unseen during the training of the multispeaker model. We use person-specific audiovisual and renderer modules and compare NEUTART's results with the aforementioned methods. We use a variety of metrics in our comparisons, both statistical and perceptual.

For evaluating the audio, we use the average Mel Cepstral Distance (MCD), a statistical measure that penalizes the deviation from the ground truth spectrogram. For a predicted spectrogram $\mathbf{F}$ is:

$$MCD = 10 \log_{10} ||\hat{\mathbf{F}} - \mathbf{F}||_2^2 \tag{6.1}$$

where $\hat{\mathbf{F}}$ would be the ground truth spectrogram. We also perceptually evaluate the intelligibility of the audio using the (audio) character error rate (ACER) from the Wav2Vec2 speech recognition model (Baevski et al. 2020). In general, *Character Error Rate* (CER) is defined as the ratio of character changes that need to happen to the transcription so that it matches the reference (Jurafsky and Martin 2021):

$$CER = \frac{\#\text{substitutions} + \#\text{insertions} + \#\text{deletions}}{\text{Total characters in the reference}} \tag{6.2}$$

ID		MCD (dB)	ACER (%)	LMD	LMVE	FID	VCER (%)	VER (%)
	Gold	-	[6.06]	-	-	-	[84.45]	[75.96]
	W2L	44.21	12.94	1.3125	0.3238	**18.36**	76.78	**68.30**
38F	ST	44.21	12.94	14.3464	0.4238	221.49	79.82	73.33
	VRT	44.21	12.94	1.6474	0.3150	37.32	80.76	75.48
	Ours	**43.41**	**10.94**	**1.1813**	**0.2889**	38.14	**74.70**	68.78
	Gold	-	[29.64]	-	-	-	[88.68]	[78.52]
	W2L	42.58	**25.01**	**1.0609**	**0.2805**	**18.45**	82.21	73.81
42M	ST	42.58	**25.01**	7.0019	0.4010	167.47	80.42	73.09
	VRT	42.58	**25.01**	1.5036	0.2753	25.34	78.97	71.38
	Ours	**42.36**	32.50	1.2073	0.2867	22.91	**78.64**	**71.15**
	Gold	-	[7.36]	-	-	-	[88.53]	[79.42]
	W2L	**43.50**	18.07	1.9305	0.4479	**17.73**	87.62	81.75
49F	ST	**43.50**	18.07	5.3592	0.5820	139.10	83.79	77.41
	VRT	**43.50**	18.07	**1.9215**	0.4347	29.47	84.76	78.15
	Ours	43.64	**16.93**	1.9576	**0.4132**	25.06	**76.88**	**72.02**
	Gold	-	[14.35]	-	-	-	[87.22]	[77.97]
	W2L	43.43	**18.67**	**1.4346**	0.3507	**18.18**	82.20	74.62
Mean	ST	43.43	**18.67**	8.9025	0.4689	176.02	81.34	74.61
	VRT	43.43	**18.67**	1.6908	0.3417	30.71	81.50	75.00
	Ours	**43.14**	20.12	1.449	**0.3296**	28.70	**76.74**	**70.65**

Table 6.2: Generation quality metrics on 3 random, unseen subjects from TCD-TIMIT. We should note that the audio evaluation of 42M does not seem comparable to the other subjects, or with the metrics extracted from the multispeaker model, leading us to believe that it is an outlier of the dataset. Nevertheless, our method performs very well in terms of lip landmark metrics and FID. Wav2Lip may have a lower FID, but it performs significantly worse than other methods in terms of human evaluation, due to the visible bounding box in the subject's mouth. Finally, our method is consistently superior when it comes to lipreading.

Method	MCD (dB)	ACER (%)
Gold	-	[16.21]
FastSpeech 2	**40.13**	27.04
Ours	40.24	**24.85**

Table 6.3: The positive impact of multimodality on audio generation. The audio produced from NEUTART is more intelligible than a plain TTS model of the same architecture, without any significant compromise on the spectrogram quality, indicating the effectiveness of visual supervision for speech. We used the multispeaker models for audio sampling in this experiment, in order to benefit from the abundance of samples.

	LMD	LMVE	FID	VCER	VER
Wav2Lip	1.6516	0.3314	**24.18**	86.67	81.71
Ours	**1.4490**	**0.3296**	28.70	**76.74**	**70.65**

Table 6.4: We fine-tuned Wav2Lip on TCD-TIMIT in order to conduct fairer comparisons with NEUTART, so that both compared models have been trained on the same data. Our objective experiments on visual generation quality show similar results to the original comparisons of Table 6.2. Notably, Wav2Lip's generalization capability deteriorated after finetuning.

The visual articulation quality is statistically assessed by the average Lip Landmark Distance (LMD) and the average Lip Landmark Velocity Error (LLVE) between the predicted video and the original. The landmarks are extracted with Google's MediaPipe framework (Kartynnik et al. 2019), to avoid any bias towards our model, had we used the same landmark detector as the one during data preprocessing (Bulat and Tzimiropoulos 2017).

Moreover, following Filntisis et al. (2023), we perform perceptual visual evaluation by cropping the mouth videos and using the pretrained AV-HuBERT (Shi et al. 2022a; Shi et al. 2022b) as a lipreader. The metrics we use are the (visual) character error rate (VCER) and the viseme error rate (VER), after mapping the phonemes into visemes using a predefined dictionary. Finally, the photorealism is evaluated with Fréchet Inception Distance (FID) (Heusel et al. 2017), a popular measure for evaluating generated images.

Note that calculating some of the aforementioned metrics requires the compared sequences to be of the same length, therefore we align them using DTW (Sakoe and Chiba 1978), a method for aligning timeseries by removing the mismatch caused by local temporal shifts. We present the evaluation results in Table 6.2. All metrics indicate better performance when their value is lower, since all of them are types of error measures.

The results indicate that NEUTART can produce articulate and coherent talking heads. We also compare NEUTART's performance on audio synthesis compared to the equivalent TTS system, which would be the FastSpeech 2 model that we used in order to drive the lip-sync methods. We present the pure audio synthesis metrics in Table 6.3. The results indicate that including a visual visual supervision can improve the generated audio's quality, especially in terms of intelligibility, as the lower audio CER suggests. This supports the effectiveness of multitask learning and showcases that visual supervision is beneficial for audio synthesis.

We should note that a crucial distinction between our architecture and the compared few-shot methods is that they use a reference face for inference. However, they have not been trained on TCD-TIMIT, which begs the question whether they would outperform NEUTART had they been trained on the same dataset, despite their few-shot design. In order to address that discrepancy of training data, we repeated the objective evaluation comparisons after fine-tuning Wav2Lip on TCD-TIMIT. Unfortunately, the other two methods have not open-sourced their training code and setup. We present the additional comparisons in Table 6.4.

We observe the same trend as in Table 6.2, with our method prevailing in most metrics. In fact,

	SadTalker	VideoRetalking	Wav2Lip	FastSpeech 2 (audio only)
NEUTART	**66** / 54 **55.0%** / 45.0%	**74** / 46 **61.7%** / 38.3%	**87** / 33 **72.5%** / 27.5%	**59** / 37 **61.5%** / 38.5%

Table 6.5: User study results with A/B preference scheme. The results show that NEUTART (left) was preferred A times, while the competing method was preferred B times, with a total of $A+B$ pairs assessed. The corresponding percentage is also shown. Users consistently judged our method as more realistic than the competing ones, both in terms of audio and video generation.

Wav2Lip's visual performance deteriorated after fine-tuning, which can be attributed to the fact that its original model was trained on the much larger LRS2 dataset (Afouras et al. 2018a) using ground truth audios of in-the-wild quality. Fine-tuning on the less massive lab-recorder TCD-TIMIT, in addition to being driven by synthetic audios, might have impacted its generalization capability due to the mismatch of driving audios between training and inference.

6.2.3 Subjective evaluation

We also conducted a user study, comparing our method against FastSpeech 2 in terms of audio realism, as well as all the aforementioned audio-driven methods in terms of audiovisual realism. To do this, we first generated a set of unseen phonetically rich text transcriptions using an appropriate prompt to a large language model. These sentences were used to generate the samples for each method, using two randomly chosen subjects (one male and one female speaker).

For the study, we adopted a preference test design. For each audio-based question users were presented with two audio files and were asked to select the one that sounds more realistic. For the audiovisual part, two synthetic videos where presented and the users were asked again to select the one that they find more realistic. Note that in each question we provided users with the transcription of the audio or video. In the case of video we also provided users with an image of the synthetically generated person from the original footage. Both order and position randomization were adopted, so the selection of the other method and the videos' order in the pair was randomized in each question. Each user answered a total of 4 audio-based questions and 15 video-based questions (5 questions for each audiovisual pair: NEUTART against some another method). A total of 21 users completed the questionnaire and the results can be seen in Table 6.5.

We see that our method is consistently perceived as more realistic by independent users. SadTalker is evaluated as the second best method, then follows VideoReTalking. We also revisit the effectiveness of multitask learning for TTS, by showing that the audiovisual model generates speech that is preferable to a plain TTS system's output. The first sample from each of the two subjects is presented in Figure 6.3, where we annotate key shortcomings of previous methods.

6.2.4 Ablation study

In order to study the effects of each additional visual loss on the talking face generation, we performed an ablation study on the TCD-TIMIT dataset. The audiovisual generator's goal is to synthesize 3D faces as accurately as possible, so we evaluated the generation in terms of audio and 3D reconstruction, with visual metrics computed from images of 3D reconstructions. We present the results in Table 6.6. We chose to perform the ablation on the multispeaker audiovisual module, observing its convergence and generation results, due to the abundance of samples. Most metrics are lower when the model is trained using all visual losses. Having validated the effectiveness of our losses, we used them for the person-specific fine-tuning.

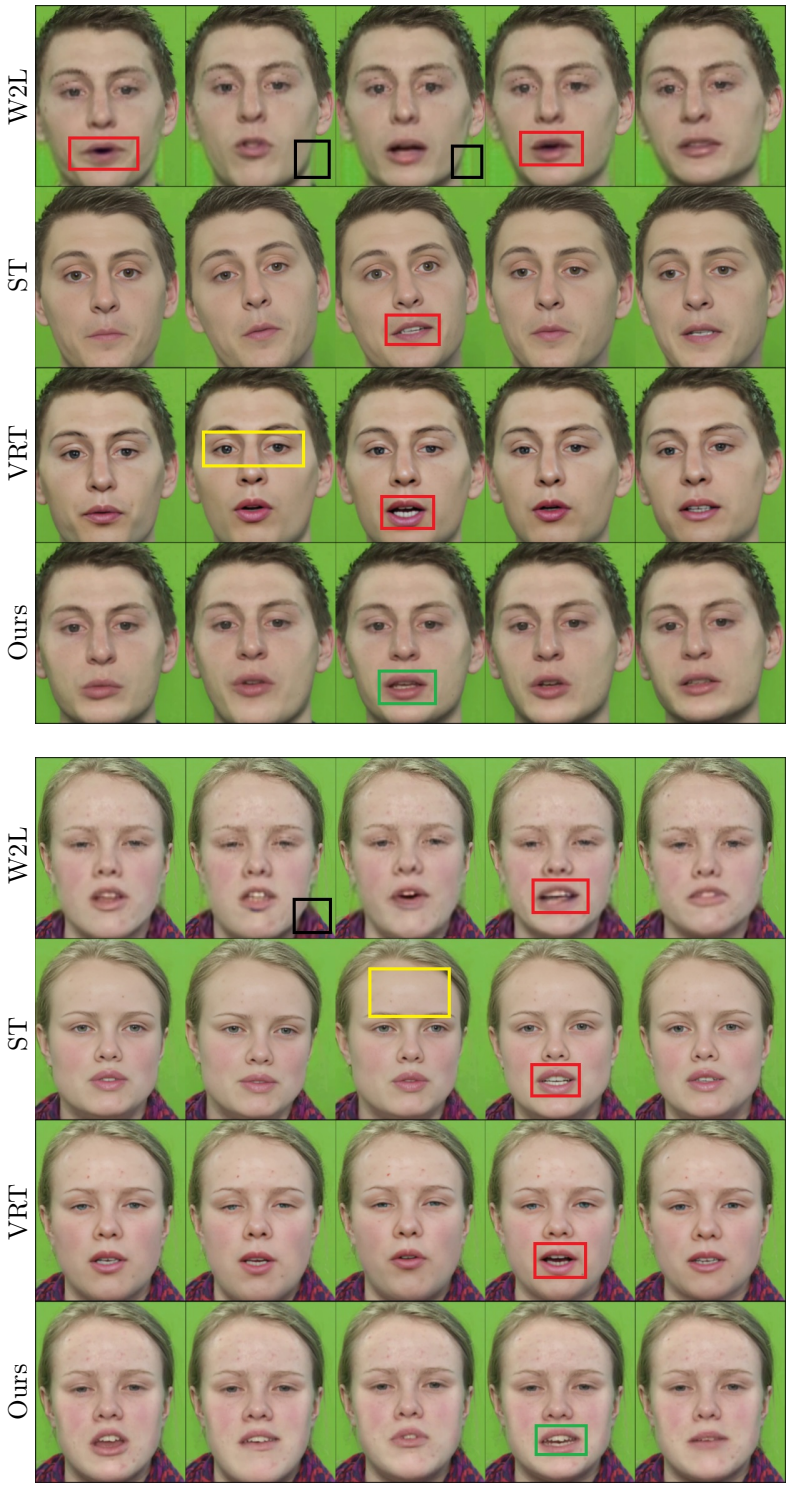


Figure 6.3: Comparison of our method against previous ones in lab conditions. Wav2Lip has poor resolution, and shows a bounding box artifact around the mouth (highlighted with black boxes). SadTalker and VideoReTalking produce frames at a much better resolution, since they use a face enhancing network, however this enhancement produces artifacts that alter the person's identity. For instance, the female subject's skin is oversmoothed, while the male subject's eye color is changed (highlighted with yellow boxes). All three of them produce some frames with unrealistic mouth interior, and seem to alter the subject's teeth (highlighted with red boxes). Our method generates frames with well-formed lips, teeth and mouth interior, without creating any uncanny effects in the subject's face.

$\mathcal{L}_{lip}$	$\mathcal{L}_{grad}$	$\mathcal{L}_{flow}$	MCD (dB)	ACER (%)	LMD	LMVE	VCER (%)	VER (%)
✗	✗	✓	41.98	**21.91**	0.5053	0.3502	82.40	77.90
✗	✓	✗	41.91	22.88	0.6856	0.3602	85.66	80.00
✗	✓	✓	**40.31**	25.05	**0.4318**	**0.3261**	77.15	70.99
✓	✓	✓	**40.31**	25.40	0.5063	0.4203	**77.05**	**70.66**

Table 6.6: Ablation study on the effectiveness of visual losses. Apart from the lipreading loss which increases the articulation realism, the usage of gradient and flow losses ensures more accurate landmark prediction. As a result, we include them for training the audiovisual module.

	SadTalker	VideoRetalking	Wav2Lip
NEUTART	**135** / 21 **86.54%** / 13.46%	65 / **91** 41.67% / **58.33%**	**107** / 49 **68.59%** / 31.41%

Table 6.7: User study results on in-the-wild videos with A/B preference scheme. The results show that NEUTART (left) was preferred A times, while the competing method was preferred B times, with a total of $A + B$ pairs assessed. The corresponding percentage is also shown. Even under in-the-wild conditions, our method yields very promising results. Users significantly preferred NEUTART over Wav2Lip and SadTalker.

6.2.5 In-the-wild experiments

We also experimented with in-the-wild videos and conducted an additional user study with 25 participants. This time, we evaluated only visual realism, due to the aforementioned limitations of audio synthesis from poor recordings. We used 2 subjects from the HDTF dataset, following the same protocol as before. We present the user study results in Table 6.7, and sample frames from each method in Figure 6.4.

It is obvious that even under in-the-wild conditions, our method yields a very promising perceived quality. Users significantly preferred NEUTART over Wav2Lip and SadTalker, with a binomial test p-value which is less than 10^{-5} in both cases. While VideoReTalking was preferred to NEUTART, this result was not so statistically significant, with the corresponding value p>0.1.

We note that experimenting with in-the-wild videos is particularly challenging for two main reasons. First of all, they are not paired with text transcriptions, which forces us to use the state-of-the-art speech recognition system named Whisper (Radford et al. 2023) in order to transcribe the speech. The Whisper Large V2 model has an average word error rate of 12.8% (see Equation 6.2 for the similar character error rate), which introduces considerable noise in the text-to-audio alignment. Secondly, they are not guaranteed to offer enough phonetic variation, causing artifacts in some challenging phonetic sequences. Thus, the audio in samples generated from those datasets is of poorer quality compared to lab-recorded datasets. Despite that, our experiments show that the audio and visual streams are consistent, with realistic lip articulation.

A qualitative conclusion that can be drawn after all the above experiments is that NEUTART is able to generate videos that achieve both audio and visual realism, when trained on data of appropriate quality. In contrast, the compared few-shot methods sacrifice some aspect of generation quality in order to be able to work on unconstrained scenarios. Nevertheless, the creation of a high-quality virtual assistant or animated avatar, either for academic or commercial purposes, would require careful data collection. In that scenario, our method would be able to show it full potential, similarly to the TCD-TIMIT results.

Figure 6.4: Comparison of our method against previous ones, with samples from two HDTF subjects. The compared methods display the same shortcomings, since they do not really differentiate between lab or in-the-wild conditions. Our method's visual generation is still very realistic and well-synced with the audio, despite the audio not being as intelligible compared to training in lab recordings.

Chapter 7

Conclusion

Contents

7.1 Summary

In this Diploma Thesis, we delved into the area of speech synthesis and talking face generation. In the beginning, we presented the some preliminaries from the area of deep learning, especially sequential architectures. Audiovisual speech synthesis, as the name suggests, pertains to the modalities of audio and vision. We first visited each modality separately by exploring the tasks of text-to-speech synthesis and 3D head modeling. Then, we moved on to the actual area of audiovisual speech synthesis, presenting the mechanisms and the latest related work.

After presenting all the above bibliography in Chapters 1-4, we moved on the experimental part. We developed a text-driven audiovisual talking face model, named NEUTART, that generates photorealistic videos. Our model approaches talking face generation as a bimodal task, with a transformer-based encoder-decoder architecture simultaneously predicting both the mel spectrogram and the 3D face parameters from a phoneme sequence. This allows for better capture of the audiovisual correlation, which benefits both modalities, and also alleviates the redundancy of extracting features from generated speech in order to create a video. The predicted 3D facial reconstructions are used as conditional input to a GAN-based neural renderer.

The model is trained on accurate 3D facial reconstructions, employing many perceptual losses such as a lipreading loss for visual supervision, which benefit the overall realism of the synthetic results. Overall, our experiments comparing NEUTART with recent state-of-the-art models reveal the following key takeaways:

- The objective and subjective evaluation shows that our model's samples are more realistic than previous few-shot approaches that do not adopt an audiovisual representation for visual speech, especially in terms of lip articulation.
- The ablation study reveals the effectiveness of visual losses in overall generation quality.
- We also showcase that treating speech synthesis as a bimodal task, by including visual supervision, enhances audio intelligibility compared to a plain TTS system.

7.2 Future Work

Towards further improving or expanding our work, we briefly highlight a few aspects that can be examined.

- **Robustness:** Our model can produce exceptional results when trained in lab-recorded datasets, thus it is tailored for creating human avatars for virtual assistants. Nevertheless, the machine learning community in the area of talking face generation is keen on exploring in-the-wild scenarios. As we have already discussed, NEUTART is lacks robustness when fine-tuned on subjects with poorer quality audio and automatic transcriptions. One possible way of tackling this shortcoming would be to research how a voice cloning architecture (Arik et al. 2018) can be adapted for audiovisual modeling.
- **Computational cost:** Performance is usually disregarded in talking face generation, especially in open-source research, whereas low computational cost and inference speed are essential for commercial applications and mobile devices. In NEUTART, the slowest component is the neural renderer, which operates needs to render the subject's head in 3D using classical rendering, then use it to predict the face in the video's pixel space. The rendering procedure can be further optimized in order to achieve faster sampling.
- **End-to-end training:** Employing end-to-end training by backpropagating the image-space error of the renderer's output back into the audiovisual submodule, may further improve the overall realism of the generated output.
- **Controllability:** It would be very interesting to pursue better guidance to the one-to-many mapping in audiovisual TTS, by using explicit labels or extracting the style from another reference video. A cutting-edge addition to the pipeline would be to control the

style via a text prompt, following the recent success of promptable image generators (Gu et al. 2023). Prompting with emotional labels, by leveraging pretrained vision-language models, has been explored in Xu et al. (2023), but to the best of our knowledge there are not any published works that are fully controllable.

7.3 Ethical Considerations

As already mentioned in Chapter 1, deep learning systems for photorealistic audiovisual speech synthesis like the one developed in this Diploma Thesis can have a very positive impact in many applications such as digital avatars, virtual assistants, accessibility tools, teleconferencing, video games, movie dubbing, and human-machine interfaces. However, at the same time, this type of technology has the risk of being misused towards unethical or malicious purposes. The misuse of such models can mislead viewers, damage the reputation of subjects, and cause acute mistrust in digital media. It may also be used produce harmful videos of individuals without their consent, raising concerns related to human rights. We refer to the following articles about an extensive discussion of these issues: Chesney and Citron (2019), Diakopoulos and Johnson (2021), and Yadlin-Segal and Oppenheim (2021).

We believe that researchers and engineers working in the relevant fields need to be mindful of these ethical concerns and contribute to raising public awareness about the capabilities of such AI systems, improving the public's media literacy. Other countermeasures include contributing in the development of state-of-the-art systems that detect deepfake videos (Zhang 2022; Masood et al. 2023). In our work, generated videos are always presented as synthetic, either explicitly or implicitly (when clearly implied by the context), and encourage other researchers and users to follow this practice. Notably, the code that was developed to train and sample from our audiovisual model, was released under an ethical license (Milis 2023). The license permits free usage of the software provided that it is used responsibly, adhering to ethical guidelines. Namely:

1. Respecting human rights, privacy, and dignity.
2. Strictly refraining from the promotion of hate speech, discrimination, violence, or any other form of harm.
3. Strictly refraining from the creation of deepfake content with any type of malicious intent.

We strongly believe in fruitful dialogue between the research community and the public, and anticipate that artificial intelligence continues to make progress towards pushing the boundaries of knowledge and producing beneficial applications for humanity.

Bibliography

Abdelaziz, Ahmed Hussen, Kumar, Anushree Prasanna, Seivwright, Chloe, Fanelli, Gabriele, Binder, Justin, Stylianou, Yannis, and Kajareker, Sachin (2021). "Audiovisual speech synthesis using tacotron2". In: *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 503–511 (cit. on pp. 8, 9, 16, 59, 81).

Abrevaya, Victoria Fernandez, Wuhrer, Stefanie, and Boyer, Edmond (2018). "Multilinear Autoencoder for 3D Face Model Learning". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9. DOI: 10.1109/WACV.2018.00007 (cit. on p. 50).

Afouras, T., Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2018a). "Deep Audio-Visual Speech Recognition". In: *arXiv:1809.02108* (cit. on p. 84).

Afouras, Triantafyllos, Chung, Joon Son, and Zisserman, Andrew (2018b). "LRS3-TED: a large-scale dataset for visual speech recognition". In: *arXiv preprint arXiv:1809.00496* (cit. on p. 50).

Agarwal, Madhav, Gupta, Anchit, Mukhopadhyay, Rudrabha, Namboodiri, Vinay P, and Jawahar, CV (2022). "Compressing Video Calls using Synthetic Talking Heads". In: *arXiv preprint arXiv:2210.03692* (cit. on p. 54).

Arik, Sercan, Chen, Jitong, Peng, Kainan, Ping, Wei, and Zhou, Yanqi (2018). "Neural Voice Cloning with a Few Samples". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. (cit. on p. 90).

Austin, Jacob, Johnson, Daniel D, Ho, Jonathan, Tarlow, Daniel, and Berg, Rianne van den (2021). "Structured denoising diffusion models in discrete state-spaces". In: *Advances in Neural Information Processing Systems* 34, pp. 17981–17993 (cit. on p. 42).

Ba, Jimmy Lei, Kiros, Jamie Ryan, and Hinton, Geoffrey E (2016). "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (cit. on p. 41).

Badlani, Rohan, Łańcucki, Adrian, Shih, Kevin J, Valle, Rafael, Ping, Wei, and Catanzaro, Bryan (2022). "One TTS alignment to rule them all". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6092–6096 (cit. on p. 40).

Baevski, Alexei, Zhou, Yuhao, Mohamed, Abdelrahman, and Auli, Michael (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in neural information processing systems* 33, pp. 12449–12460 (cit. on p. 81).

Bao, Linchao, Zhang, Haoxian, Qian, Yue, Xue, Tangli, Chen, Changhai, Zhe, Xuefei, and Kang, Di (2023). "Learning Audio-Driven Viseme Dynamics for 3D Face Animation". In: *arXiv preprint arXiv:2301.06059* (cit. on p. 55).

Bigioi, Dan, Basak, Shubhajit, Jordan, Hugh, McDonnell, Rachel, and Corcoran, Peter (2023). "Speech Driven Video Editing via an Audio-Conditioned Diffusion Model". In: *arXiv preprint arXiv:2301.04474* (cit. on p. 57).

Bishop, Christopher M and Nasrabadi, Nasser M (2006). *Pattern recognition and machine learning*. Vol. 4. 4. Springer (cit. on pp. 1, 25).

Blanz, Volker and Vetter, Thomas (1999). "A morphable model for the synthesis of 3D faces". In: *International Conference on Computer Graphics and Interactive Techniques* (cit. on pp. 5, 45, 46, 48).

Boersma, Paul and Weenink, David (2009). *Praat: doing phonetics by computer (Version 5.1.13)*. URL: http://www.praat.org (cit. on p. 38).

Booth, James, Roussos, Anastasios, Ponniah, Allan, Dunaway, David, and Zafeiriou, Stefanos (Apr. 2018). "Large Scale 3D Morphable Models". In: *Int. J. Comput. Vision* 126.2–4, pp. 233–254. ISSN: 0920-5691. DOI: 10.1007/s11263-017-1009-7 (cit. on pp. 6, 46, 47, 50).

Bulat, Adrian and Tzimiropoulos, Georgios (2017). "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1021–1030 (cit. on pp. 11, 66, 83).

Cao, Houwei, Cooper, David G, Keutmann, Michael K, Gur, Ruben C, Nenkova, Ani, and Verma, Ragini (2014). "Crema-d: Crowd-sourced emotional multimodal actors dataset". In: *IEEE transactions on affective computing* 5.4, pp. 377–390 (cit. on p. 78).

Chalamandaris, Aimilios, Karabetsos, Sotiris, Tsiakoulis, Pirros, and Raptis, Spyros (2010). "A unit selection text-to-speech synthesis system optimized for use with screen readers". In: *IEEE Transactions on Consumer Electronics* 56.3, pp. 1890–1897 (cit. on p. 40).

Chatziagapi, Aggelina and Samaras, Dimitris (2023). "AVFace: Towards Detailed Audio-Visual 4D Face Reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16878–16889 (cit. on pp. 6, 50).

Chen, Lele, Cui, Guofeng, Kou, Ziyi, Zheng, Haitian, and Xu, Chenliang (2020a). "What comprises a good talking-head video generation?" In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (cit. on pp. 16, 61, 81).

Chen, Nanxin, Zhang, Yu, Zen, Heiga, Weiss, Ron J, Norouzi, Mohammad, and Chan, William (2020b). "Wavegrad: Estimating gradients for waveform generation". In: *arXiv preprint arXiv:2009.00713* (cit. on p. 41).

Chen, Xin, Jiang, Biao, Liu, Wen, Huang, Zilong, Fu, Bin, Chen, Tao, Yu, Jingyi, and Yu, Gang (2022). "Executing your Commands via Motion Diffusion in Latent Space". In: *arXiv preprint arXiv:2212.04048* (cit. on p. 32).

Cheng, Kun, Cun, Xiaodong, Zhang, Yong, Xia, Menghan, Yin, Fei, Zhu, Mingrui, Wang, Xuan, Wang, Jue, and Wang, Nannan (2022). "Videoretalking: Audio-based lip synchronization for talking head video editing in the wild". In: *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9 (cit. on pp. 8, 17, 56, 57, 64, 81).

Chesney, Bobby and Citron, Danielle (2019). "Deep fakes: A looming challenge for privacy, democracy, and national security". In: *Calif. L. Rev.* 107, p. 1753 (cit. on pp. 21, 91).

Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). "Lip Reading Sentences in the Wild". In: *IEEE Conference on Computer Vision and Pattern Recognition* (cit. on p. 78).

Chung, Joon Son and Zisserman, Andrew (2017). "Out of time: automated lip sync in the wild". In: *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*. Springer, pp. 251–263 (cit. on p. 64).

Coker, Cecil H (1976). "A model of articulatory dynamics and control". In: *Proceedings of the IEEE* 64.4, pp. 452–460 (cit. on p. 39).

Cootes, Timothy F., Edwards, Gareth J., and Taylor, Christopher J (2001). "Active appearance models". In: *IEEE Transactions on pattern analysis and machine intelligence* 23.6, pp. 681–685 (cit. on p. 54).

Craig, John (2021). *Introduction to Robotics, Global Edition*. Pearson Education (cit. on p. 69).

Cudeiro, Daniel, Bolkart, Timo, Laidlaw, Cassidy, Ranjan, Anurag, and Black, Michael J (2019). "Capture, learning, and synthesis of 3D speaking styles". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10101–10111 (cit. on pp. 7, 55).

Daněček, Radek, Black, Michael J, and Bolkart, Timo (2022). "EMOCA: Emotion driven monocular face capture and animation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20322 (cit. on pp. 6, 49).

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (cit. on p. 42).

Diakopoulos, Nicholas and Johnson, Deborah (2021). "Anticipating and addressing the ethical implications of deepfakes in the context of elections". In: *New Media & Society* 23.7, pp. 2072–2098 (cit. on pp. 21, 91).

Doukas, Michail Christos, Koujan, Mohammad Rami, Sharmanska, Viktoriia, Roussos, Anastasios, and Zafeiriou, Stefanos (2021a). "Head2Head++: Deep Facial Attributes Re-Targeting". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3.1, pp. 31–43. DOI: 10.1109/TBIOM.2021.3049576 (cit. on p. 56).

– (2021b). "Head2head++: Deep facial attributes re-targeting". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3.1, pp. 31–43 (cit. on pp. 8, 14, 66, 72).

Doukas, Michail Christos, Zafeiriou, Stefanos, and Sharmanska, Viktoriia (2021c). "Headgan: One-shot neural head synthesis and editing". In: *Proceedings of the IEEE/CVF International conference on Computer Vision*, pp. 14398–14407 (cit. on p. 56).

Du, Chenpng, Chen, Qi, He, Tianyu, Tan, Xu, Chen, Xie, Yu, Kai, Zhao, Sheng, and Bian, Jiang (2023). "DAE-Talker: High Fidelity Speech-Driven Talking Face Generation with Diffusion Autoencoder". In: *arXiv preprint arXiv:2303.17550* (cit. on p. 57).

Dutoit, Thierry (1997). "High-quality text-to-speech synthesis: An overview". In: *Journal Of Electrical And Electronics Engineering Australia* 17.1, pp. 25–36 (cit. on p. 40).

Egger, Bernhard, Smith, William AP, Tewari, Ayush, Wuhrer, Stefanie, Zollhoefer, Michael, Beeler, Thabo, Bernard, Florian, Bolkart, Timo, Kortylewski, Adam, Romdhani, Sami, et al. (2020). "3d morphable face models—past, present, and future". In: *ACM Transactions on Graphics (TOG)* 39.5, pp. 1–38 (cit. on p. 45).

Ezzat, Tony and Poggio, Tomaso A. (2000). "Visual Speech Synthesis by Morphing Visemes". In: *International Journal of Computer Vision* 38, pp. 45–57 (cit. on p. 58).

Fan, Yingruo, Lin, Zhaojiang, Saito, Jun, Wang, Wenping, and Komura, Taku (2022). "Faceformer: Speech-driven 3d facial animation with transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18770–18780 (cit. on pp. 7, 55).

Feng, Yao, Feng, Haiwen, Black, Michael J, and Bolkart, Timo (2021). "Learning an animatable detailed 3d face model from in-the-wild images". In: *ACM Transactions on Graphics (ToG)* 40.4, pp. 1–13 (cit. on pp. 6, 48, 56, 75).

Filntisis, Panagiotis P, Retsinas, George, Paraperas-Papantoniou, Foivos, Katsamanis, Athanasios, Roussos, Anastasios, and Maragos, Petros (2023). "SPECTRE: Visual Speech-Informed Perceptual 3D Facial Expression Reconstruction From Videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5744–5754 (cit. on pp. 6, 7, 9, 14, 49, 50, 64, 66, 71, 83).

Filntisis, Panagiotis P., Katsamanis, Athanasios, and Maragos, Petros (2017a). "Photorealistic adaptation and interpolation of facial expressions using HMMS and AAMS for audio-visual speech synthesis". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 2941–2945 (cit. on p. 58).

Filntisis, Panagiotis P., Katsamanis, Athanasios, Tsiakoulis, Pirros, and Maragos, Petros (2017b). "Video-realistic expressive audio-visual speech synthesis for the Greek language". In: *Speech Commun.* 95, pp. 137–152 (cit. on p. 58).

Ghosh, Partha, Gupta, Pravir Singh, Uziel, Roy, Ranjan, Anurag, Black, Michael J, and Bolkart, Timo (2020). "GIF: Generative interpretable faces". In: *2020 International Conference on 3D Vision (3DV)*. IEEE, pp. 868–878 (cit. on p. 30).

Goodfellow, Ian (2016). "Nips 2016 tutorial: Generative adversarial networks". In: *arXiv preprint arXiv:1701.00160* (cit. on pp. 3, 29, 30).

Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press (cit. on pp. 1, 24).

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc. (cit. on pp. 29, 73).

Groth, Colin, Tauscher, Jan-Philipp, Castillo, Susana, and Magnor, Marcus (2020). "Altering the Conveyed Facial Emotion Through Automatic Reenactment of Video Portraits". In: *Proceedings of the International Conference on Computer Animation and Social Agents (CASA)*. Vol. 1300, pp. 128–135 (cit. on p. 56).

Gu, Jindong, Han, Zhen, Chen, Shuo, Beirami, Ahmad, He, Bailan, Zhang, Gengyuan, Liao, Ruotong, Qin, Yao, Tresp, Volker, and Torr, Philip (2023). "A systematic survey of prompt engineering on vision-language foundation models". In: *arXiv preprint arXiv:2307.12980* (cit. on p. 91).

Guan, Jiazhi, Zhang, Zhanwang, Zhou, Hang, Hu, Tianshu, Wang, Kaisiyuan, He, Dongliang, Feng, Haocheng, Liu, Jingtuo, Ding, Errui, Liu, Ziwei, et al. (2023). "StyleSync: High-Fidelity Generalized and Personalized Lip Sync in Style-based Generator". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1505–1515 (cit. on p. 64).

Gulati, Anmol, Qin, James, Chiu, Chung-Cheng, Parmar, Niki, Zhang, Yu, Yu, Jiahui, Han, Wei, Wang, Shibo, Zhang, Zhengdong, Wu, Yonghui, et al. (2020). "Conformer: Convolution-augmented transformer for speech recognition". In: *arXiv preprint arXiv:2005.08100* (cit. on p. 57).

Guo, Zhifang, Leng, Yichong, Wu, Yihan, Zhao, Sheng, and Tan, Xu (2023). "Prompttts: Controllable Text-To-Speech With Text Descriptions". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on p. 42).

Hannun, Awni, Case, Carl, Casper, Jared, Catanzaro, Bryan, Diamos, Greg, Elsen, Erich, Prenger, Ryan, Satheesh, Sanjeev, Sengupta, Shubho, Coates, Adam, et al. (2014). "Deep speech: Scaling up end-to-end speech recognition". In: *arXiv preprint arXiv:1412.5567* (cit. on pp. 55, 56).

Harte, Naomi and Gillen, Eoin (2015). "TCD-TIMIT: An audio-visual corpus of continuous speech". In: *IEEE Transactions on Multimedia* 17.5, pp. 603–615 (cit. on pp. 4, 16, 38, 78, 79).

Hartley, R and Zisserman, A (2003). "Multiple view geometry in computer". In: *Vision, 2nd ed., New York: Cambridge* 2, p. 5 (cit. on p. 45).

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (cit. on pp. 27, 48).

Heusel, Martin, Ramsauer, Hubert, Unterthiner, Thomas, Nessler, Bernhard, and Hochreiter, Sepp (2017). "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30 (cit. on pp. 60, 83).

Ho, Jonathan, Jain, Ajay, and Abbeel, Pieter (2020). "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33, pp. 6840–6851 (cit. on pp. 3, 30, 31).

Ho, Jonathan, Salimans, Tim, Gritsenko, Alexey, Chan, William, Norouzi, Mohammad, and Fleet, David J (2022). "Video diffusion models". In: *arXiv preprint arXiv:2204.03458* (cit. on p. 32).

Hochreiter, Sepp and Schmidhuber, Jürgen (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780 (cit. on p. 27).

Hussen Abdelaziz, Ahmed, Theobald, Barry-John, Dixon, Paul, Knothe, Reinhard, Apostoloff, Nicholas, and Kajareker, Sachin (2020). "Modality dropout for improved performance-driven talking faces". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 378–386 (cit. on pp. 13, 71).

Hwang, Geumbyeol, Hong, Sunwon, Lee, Seunghyun, Park, Sungwoo, and Chae, Gyeongsu (2023). "DisCoHead: Audio-and-Video-Driven Talking Head Generation by Disentangled Control of Head Pose and Facial Expressions". In: *arXiv preprint arXiv:2303.07697* (cit. on p. 57).

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A (2017). "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134 (cit. on p. 64).

Ito, Keith and Johnson, Linda (2017). *The LJ Speech Dataset*. `https://keithito.com/LJ-Speech-Dataset/` (cit. on pp. 16, 78).

Jackson, Aaron S, Bulat, Adrian, Argyriou, Vasileios, and Tzimiropoulos, Georgios (2017). "Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression". In: *International Conference on Computer Vision* (cit. on pp. 6, 50).

Ji, Xinya, Zhou, Hang, Wang, Kaisiyuan, Wu, Qianyi, Wu, Wayne, Xu, Feng, and Cao, Xun (2022). "EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model". In: *ACM SIGGRAPH 2022 Conference Proceedings*. SIGGRAPH '22. ISBN: 9781450393379. DOI: 10.1145/3528233.3530745 (cit. on p. 58).

Jurafsky, Dan and Martin, James H. (2021). *Speech and language processing, 3rd Edition* (cit. on p. 81).

Kalchbrenner, Nal, Elsen, Erich, Simonyan, Karen, Noury, Seb, Casagrande, Norman, Lockhart, Edward, Stimberg, Florian, Oord, Aaron, Dieleman, Sander, and Kavukcuoglu, Koray (2018). "Efficient neural audio synthesis". In: *International Conference on Machine Learning*. PMLR, pp. 2410–2419 (cit. on pp. 8, 59).

Kang, Minguk, Zhu, Jun-Yan, Zhang, Richard, Park, Jaesik, Shechtman, Eli, Paris, Sylvain, and Park, Taesung (2023). "Scaling up gans for text-to-image synthesis". In: *arXiv preprint arXiv:2303.05511* (cit. on p. 30).

Karras, Tero, Laine, Samuli, and Aila, Timo (2019). "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410 (cit. on p. 30).

Karras, Tero, Laine, Samuli, Aittala, Miika, Hellsten, Janne, Lehtinen, Jaakko, and Aila, Timo (2020). "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119 (cit. on pp. 3, 30).

Kartynnik, Yury, Ablavatski, Artsiom, Grishchenko, Ivan, and Grundmann, Matthias (2019). "Real-time facial surface geometry from monocular video on mobile GPUs". In: *arXiv preprint arXiv:1907.06724* (cit. on p. 83).

Kim, Hyeongwoo, Elgharib, Mohamed, Zollhöfer, Michael, Seidel, Hans-Peter, Beeler, Thabo, Richardt, Christian, and Theobalt, Christian (Nov. 2019). "Neural Style-Preserving Visual Dubbing". In: *ACM Trans. Graph.* 38.6. ISSN: 0730-0301. DOI: 10.1145/3355089.3356500 (cit. on p. 56).

Kim, Hyeongwoo, Garrido, Pablo, Tewari, Ayush, Xu, Weipeng, Thies, Justus, Nießner, Matthias, Pérez, Patrick, Richardt, Christian, Zollöfer, Michael, and Theobalt, Christian (2018). "Deep Video Portraits". In: *ACM Transactions on Graphics (TOG)* 37.4, p. 163 (cit. on p. 56).

Kim, Jaehyeon, Kim, Sungwon, Kong, Jungil, and Yoon, Sungroh (2020). "Glow-tts: A generative flow for text-to-speech via monotonic alignment search". In: *Advances in Neural Information Processing Systems* 33, pp. 8067–8077 (cit. on p. 41).

Kim, Jaehyeon, Kong, Jungil, and Son, Juhee (2021a). "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech". In: *International Conference on Machine Learning*. PMLR, pp. 5530–5540 (cit. on p. 60).

Kim, Kihong, Kim, Yunho, Cho, Seokju, Seo, Junyoung, Nam, Jisu, Lee, Kychul, Kim, Seungryong, and Lee, KwangHee (2022). "DiffFace: Diffusion-based Face Swapping with Facial Guidance". In: *arXiv preprint arXiv:2212.13344* (cit. on p. 32).

Kim, Minchan, Cheon, Sung Jun, Choi, Byoung Jin, Kim, Jong Jin, and Kim, Nam Soo (2021b). "Expressive text-to-speech using style tag". In: *arXiv preprint arXiv:2104.00436* (cit. on p. 42).

Kong, Jungil, Kim, Jaehyeon, and Bae, Jaekyoung (2020a). "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis". In: *Advances in Neural Information Processing Systems* 33, pp. 17022–17033 (cit. on pp. 11, 40, 69).

Kong, Zhifeng, Ping, Wei, Huang, Jiaji, Zhao, Kexin, and Catanzaro, Bryan (2020b). "Diffwave: A versatile diffusion model for audio synthesis". In: *arXiv preprint arXiv:2009.09761* (cit. on p. 41).

Kreis, Karsten, Gao, Ruiqi, and Vahdat, Arash (2022). "Denoising Diffusion-based Generative Modeling: Foundations and Applications". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. URL: https://cvpr2022-tutorial-diffusion-models.github.io/ (cit. on p. 32).

Kumar, Rithesh, Sotelo, Jose, Kumar, Kundan, Brébisson, Alexandre de, and Bengio, Yoshua (2017). "Obamanet: Photo-realistic lip-sync from text". In: *arXiv preprint arXiv:1801.01442* (cit. on pp. 8, 58).

Lan, Chong, Wang, Yongsheng, Wang, Chengze, Song, Shirong, and Gong, Zheng (2023). "Application of ChatGPT-Based Digital Human in Animation Creation". In: *Future Internet* 15.9, p. 300 (cit. on p. 64).

Łańcucki, Adrian (2021). "Fastpitch: Parallel text-to-speech with pitch prediction". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6588–6592 (cit. on p. 68).

LeCun, Yann, Bengio, Y., and Hinton, Geoffrey (May 2015). "Deep Learning". In: *Nature* 521, pp. 436–44. DOI: 10.1038/nature14539 (cit. on pp. 1, 24, 26, 27).

Lee, Jiyoung, Chung, Joon Son, and Chung, Soo-Whan (2023). "Imaginary Voice: Face-styled Diffusion Model for Text-to-Speech". In: *arXiv preprint arXiv:2302.13700* (cit. on p. 42).

Li, Lincheng, Wang, Suzhen, Zhang, Zhimeng, Ding, Yu, Zheng, Yixing, Yu, Xin, and Fan, Changjie (2021). "Write-a-speaker: Text-based emotional and rhythmic talking-head generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 3, pp. 1911–1920 (cit. on p. 59).

Li, Tianye, Bolkart, Timo, Black, Michael J, Li, Hao, and Romero, Javier (2017). "Learning a model of facial shape and expression from 4D scans." In: *ACM Trans. Graph.* 36.6, pp. 194–1 (cit. on pp. 6, 47, 58, 66, 68).

Liu, Jinglin, Zhu, Zhiying, Ren, Yi, Huang, Wencan, Huai, Baoxing, Yuan, Nicholas, and Zhao, Zhou (2022). "Parallel and High-Fidelity Text-to-Lip Generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2, pp. 1738–1746 (cit. on p. 59).

Lowe, David G (2004). "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60, pp. 91–110 (cit. on p. 45).

Ma, Dabiao, Su, Zhiba, Wang, Wenxuan, and Lu, Yuhao (2020). "FPETS: fully parallel end-to-end text-to-speech system". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 8457–8463 (cit. on p. 40).

Ma, Pingchuan, Petridis, Stavros, and Pantic, Maja (2022). "Visual Speech Recognition for Multiple Languages in the Wild". In: *Nature Machine Intelligence* 4, pp. 930–939. DOI: 10.1038/s42256-022-00550-z (cit. on pp. 13, 71).

Ma, Yifeng, Wang, Suzhen, Ding, Yu, Ma, Bowen, Lv, Tangjie, Fan, Changjie, Hu, Zhipeng, Deng, Zhidong, and Yu, Xin (2023). "TalkCLIP: Talking Head Generation with Text-Guided Expressive Speaking Styles". In: *arXiv preprint arXiv:2304.00334* (cit. on p. 58).

Mao, Xudong, Li, Qing, Xie, Haoran, Lau, Raymond YK, Wang, Zhen, and Paul Smolley, Stephen (2017). "Least squares generative adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802 (cit. on pp. 14, 74).

Masood, Momina, Nawaz, Mariam, Malik, Khalid Mahmood, Javed, Ali, Irtaza, Aun, and Malik, Hafiz (2023). "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward". In: *Applied intelligence* 53.4, pp. 3974–4026 (cit. on p. 91).

McAuliffe, Michael, Socolof, Michaela, Mihuc, Sarah, Wagner, Michael, and Sonderegger, Morgan (2017). "Montreal forced aligner: Trainable text-speech alignment using kaldi." In: *Interspeech*. Vol. 2017, pp. 498–502 (cit. on pp. 11, 38, 66).

Medina, Salvador, Tome, Denis, Stoll, Carsten, Tiede, Mark, Munhall, Kevin, Hauptmann, Alex, and Matthews, Iain (2022). "Speech Driven Tongue Animation". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20374–20384. DOI: 10.1109/CVPR52688.2022.01976 (cit. on p. 59).

Milis, Georgios (2023). *Neural Text to Articulate Talk: Deep Text to Audiovisual Speech Synthesis achieving both Auditory and Photo-realism*. https://g-milis.github.io/neutart.html (cit. on pp. 33, 91).

Min, Dongchan, Lee, Dong Bok, Yang, Eunho, and Hwang, Sung Ju (2021). "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation". In: *International Conference on Machine Learning*. PMLR, pp. 7748–7759 (cit. on p. 68).

Mirza, Mehdi and Osindero, Simon (2014). "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (cit. on p. 29).

Mitsui, Kentaro, Hono, Yukiya, and Sawada, Kei (2023). "UniFLG: Unified Facial Landmark Generator from Text or Speech". In: *arXiv preprint arXiv:2302.14337* (cit. on pp. 8, 16, 59, 60, 81).

Mori, Masahiro, Macdorman, Karl F., and Kageki, Norri (2012). "The Uncanny Valley [From the Field]". In: *IEEE Robotics Autom. Mag.* 19, pp. 98–100 (cit. on p. 54).

Morise, Masanori, Yokomori, Fumiya, and Ozawa, Kenji (2016). "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications". In: *IEICE TRANSACTIONS on Information and Systems* 99.7, pp. 1877–1884 (cit. on p. 68).

Moulines, Éric and Charpentier, Francis (1989). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". In: *Speech Commun.* 9, pp. 453–467 (cit. on p. 39).

Müller, Meinard (2007). "Dynamic time warping". In: *Information retrieval for music and motion*, pp. 69–84 (cit. on p. 60).

Nair, Nithin Gopalakrishnan, Bandara, Wele Gedara Chaminda, and Patel, Vishal M (2022). "Image generation with multimodal priors using denoising diffusion probabilistic models". In: *arXiv preprint arXiv:2206.05039* (cit. on p. 32).

Nirkin, Yuval, Keller, Yosi, and Hassner, Tal (2019). "Fsgan: Subject agnostic face swapping and reenactment". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7184–7193 (cit. on pp. 9, 65).

Novaković, P, Hornak, M, Zachar, MJ, and Joncic, N (2017). "3D Digital Recording of Archaeological, Architectural and Artistic Heritage". In: *University of Ljubljana, Ljubljana* (cit. on p. 44).

Obradović, Vladimir, Rajak, Ilija, Sečujski, Milan, and Delič, Vlado (2022). "Text driven virtual speakers". In: *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 1170–1173. DOI: 10.23919/EUSIPCO55093.2022.9909813 (cit. on pp. 8, 58).

Oord, Aäron van den, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew W., and Kavukcuoglu, Koray (2016). "WaveNet: A Generative Model for Raw Audio". In: *CoRR* abs/1609.03499. arXiv: 1609.03499 (cit. on p. 40).

Oussidi, Achraf and Elhassouny, Azeddine (2018). "Deep generative models: Survey". In: *2018 International conference on intelligent systems and computer vision (ISCV)*. IEEE, pp. 1–8 (cit. on p. 29).

Paraperas Papantoniou, Foivos, Filntisis, Panagiotis P., Maragos, Petros, and Roussos, Anastasios (2022). "Neural Emotion Director: Speech-preserving semantic control of facial expressions in "in-the-wild" videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on pp. 8, 14, 56, 66, 72, 73).

Park, Kyubyong and Kim, Jongseok (2019). *g2pE*. https://github.com/Kyubyong/g2p (cit. on pp. 11, 69).

Park, Se Jin, Kim, Minsu, Hong, Joanna, Choi, Jeongsoo, and Ro, Yong Man (2022). "Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2, pp. 2062–2070 (cit. on p. 64).

Parke, Frederic I and Waters, Keith (2008). *Computer facial animation*. CRC press (cit. on p. 54).

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca, et al. (2019). "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (cit. on p. 26).

Peng, Ziqiao, Wu, Haoyu, Song, Zhenbo, Xu, Hao, Zhu, Xiangyu, Liu, Hongyan, He, Jun, and Fan, Zhaoxin (2023). "EmoTalk: Speech-driven emotional disentanglement for 3D face animation". In: *arXiv preprint arXiv:2303.11089* (cit. on p. 58).

Popov, Vadim, Vovk, Ivan, Gogoryan, Vladimir, Sadekova, Tasnima, and Kudinov, Mikhail (2021). "Grad-tts: A diffusion probabilistic model for text-to-speech". In: *International Conference on Machine Learning*. PMLR, pp. 8599–8608 (cit. on p. 41).

Prajwal, KR, Mukhopadhyay, Rudrabha, Namboodiri, Vinay P, and Jawahar, CV (2020). "A lip sync expert is all you need for speech to lip generation in the wild". In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 484–492 (cit. on pp. 8, 16, 55, 58, 64, 81).

Rabiner, Lawrence and Schafer, Ronald (2010). *Theory and Applications of Digital Speech Processing*. 1st. Prentice Hall Press (cit. on pp. 4, 36).

Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, et al. (2021). "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR, pp. 8748–8763 (cit. on pp. 30, 32).

Radford, Alec, Kim, Jong Wook, Xu, Tao, Brockman, Greg, McLeavey, Christine, and Sutskever, Ilya (2023). "Robust speech recognition via large-scale weak supervision". In: *International Conference on Machine Learning*. PMLR, pp. 28492–28518 (cit. on p. 86).

Radford, Alec, Metz, Luke, and Chintala, Soumith (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434* (cit. on pp. 3, 30, 31).

Ranjan, Anurag, Bolkart, Timo, Sanyal, Soubhik, and Black, Michael J (2018). "Generating 3D faces using convolutional mesh autoencoders". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 704–720 (cit. on p. 50).

Ren, Yi, Hu, Chenxu, Tan, Xu, Qin, Tao, Zhao, Sheng, Zhao, Zhou, and Liu, Tie-Yan (2020). "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech". In: *International Conference on Learning Representations* (cit. on pp. 11, 40, 41, 68, 79).

Ren, Yi, Ruan, Yangjun, Tan, Xu, Qin, Tao, Zhao, Sheng, Zhao, Zhou, and Liu, Tie-Yan (2019). "Fastspeech: Fast, robust and controllable text to speech". In: *Advances in neural information processing systems* 32 (cit. on p. 40).

Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, and Ommer, Björn (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695 (cit. on p. 31).

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, pp. 234–241 (cit. on pp. 32, 40, 65).

Ruan, Ludan, Ma, Yiyang, Yang, Huan, He, Huiguo, Liu, Bei, Fu, Jianlong, Yuan, Nicholas Jing, Jin, Qin, and Guo, Baining (2022). "MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation". In: *arXiv preprint arXiv:2212.09478* (cit. on p. 32).

Sagisaka, Yoshinori, Kaiki, Nobuyoshi, Iwahashi, Naoto, and Mimura, Katsuhiko (1992). "ATR μ-talk speech synthesis system". In: *Proc. 2nd International Conference on Spoken Language Processing (ICSLP 1992)*, pp. 483–486. DOI: 10.21437/ICSLP.1992-125 (cit. on pp. 24, 40).

Sakoe, Hiroaki and Chiba, Seibi (1978). "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE transactions on acoustics, speech, and signal processing* 26.1, pp. 43–49 (cit. on p. 83).

Sandler, Mark, Howard, Andrew, Zhu, Menglong, Zhmoginov, Andrey, and Chen, Liang-Chieh (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520 (cit. on p. 50).

Saragih, Jason M, Lucey, Simon, and Cohn, Jeffrey F (2011). "Real-time avatar animation from a single image". In: *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, pp. 117–124 (cit. on p. 66).

Shadle, Christine H and Damper, Robert I (2002). "Prospects for articulatory synthesis: A position paper". In: (cit. on p. 39).

Shen, Jonathan, Pang, Ruoming, Weiss, Ron J, Schuster, Mike, Jaitly, Navdeep, Yang, Zongheng, Chen, Zhifeng, Zhang, Yu, Wang, Yuxuan, Skerrv-Ryan, Rj, et al. (2018). "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 4779–4783 (cit. on pp. 8, 40, 59).

Shen, Shuai, Zhao, Wenliang, Meng, Zibin, Li, Wanhua, Zhu, Zheng, Zhou, Jie, and Lu, Jiwen (2023). "DiffTalk: Crafting Diffusion Models for Generalized Talking Head Synthesis". In: *arxiv* (cit. on p. 57).

Sheng, Changchong, Kuang, Gangyao, Bai, Liang, Hou, Chenping, Guo, Yulan, Xu, Xin, Pietikäinen, Matti, and Liu, Li (2022). "Deep Learning for Visual Speech Analysis: A Survey". In: *arXiv preprint arXiv:2205.10839* (cit. on pp. 8, 55).

Sherstinsky, Alex (2020). "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network". In: *Physica D: Nonlinear Phenomena* 404, p. 132306 (cit. on p. 27).

Shi, Bowen, Hsu, Wei-Ning, Lakhotia, Kushal, and Mohamed, Abdelrahman (2022a). "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction". In: *arXiv preprint arXiv:2201.02184* (cit. on pp. 17, 83).

Shi, Bowen, Hsu, Wei-Ning, and Mohamed, Abdelrahman (2022b). "Robust Self-Supervised Audio-Visual Speech Recognition". In: *arXiv preprint arXiv:2201.01763* (cit. on pp. 17, 83).

Shih, Kevin J, Valle, Rafael, Badlani, Rohan, Lancucki, Adrian, Ping, Wei, and Catanzaro, Bryan (2021). "RAD-TTS: Parallel flow-based TTS with robust alignment learning and diverse synthesis". In: *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models* (cit. on p. 40).

Simonyan, K and Zisserman, A (2015). "Very deep convolutional networks for large-scale image recognition". In: Computational and Biological Learning Society, pp. 1–14 (cit. on pp. 15, 74).

Sohl-Dickstein, Jascha, Weiss, Eric, Maheswaranathan, Niru, and Ganguli, Surya (2015). "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning*. PMLR, pp. 2256–2265 (cit. on pp. 3, 30, 31).

Solanki, Girish Kumar and Roussos, Anastasios (2021). "Deep Semantic Manipulation of Facial Videos". In: *arXiv preprint arXiv:2111.07902* (cit. on p. 56).

Song, Hyoung-Kyu, Woo, Sang Hoon, Lee, Junhyeok, Yang, Seungmin, Cho, Hyunjae, Lee, Youseong, Choi, Dongho, and Kim, Kang-wook (2022). "Talking Face Generation with Multilingual TTS". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21425–21430 (cit. on pp. 8, 58).

Song, Jiaming, Meng, Chenlin, and Ermon, Stefano (Oct. 2020). "Denoising Diffusion Implicit Models". In: *arXiv:2010.02502* (cit. on p. 32).

Stevens, Stanley Smith, Volkmann, John, and Newman, Edwin Broomell (1937). "A scale for the measurement of the psychological magnitude pitch". In: *The journal of the acoustical society of america* 8.3, pp. 185–190 (cit. on pp. 5, 38).

Stypułkowski, Michał, Vougioukas, Konstantinos, He, Sen, Zięba, Maciej, Petridis, Stavros, and Pantic, Maja (2023). "Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation". In: *https://arxiv.org/abs/2301.03396* (cit. on p. 56).

Tachibana, Hideyuki, Uenoyama, Katsuya, and Aihara, Shunsuke (2018). "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 4784–4788 (cit. on p. 40).

Tan, Xu, Qin, Tao, Soong, Frank, and Liu, Tie-Yan (2021). "A survey on neural speech synthesis". In: *arXiv preprint arXiv:2106.15561* (cit. on pp. 24, 40).

Tevet, Guy, Raab, Sigal, Gordon, Brian, Shafir, Yonatan, Cohen-Or, Daniel, and Bermano, Amit H (2022). "Human motion diffusion model". In: *arXiv preprint arXiv:2209.14916* (cit. on p. 32).

Theobald, Barry-John, Fagel, Sascha, Bailly, Gérard, and Elisei, Frédéric (2008). "LIPS2008: Visual speech synthesis challenge". In: *Interspeech 2008-9th Annual Conference of the International Speech Communication Association*, pp. 2310–2313 (cit. on pp. 16, 78).

Thies, Justus, Elgharib, Mohamed, Tewari, Ayush, Theobalt, Christian, and Nießner, Matthias (2020). "Neural voice puppetry: Audio-driven facial reenactment". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, pp. 716–731 (cit. on pp. 7, 55, 56).

Tokuda, Keiichi, Yoshimura, Takayoshi, Masuko, Takashi, Kobayashi, Takao, and Kitamura, Tadashi (2000). "Speech parameter generation algorithms for HMM-based speech synthesis". In: *2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100)*. Vol. 3. IEEE, pp. 1315–1318 (cit. on p. 40).

Toshpulatov, Mukhiddin, Lee, Wookey, and Lee, Suan (2023). "Talking human face generation: A survey". In: *Expert Systems with Applications*, p. 119678 (cit. on p. 54).

Tripathy, Soumya, Kannala, Juho, and Rahtu, Esa (2020). "ICface: Interpretable and Controllable Face Reenactment Using GANs". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (cit. on p. 56).

– (2021). "FACEGAN: Facial Attribute Controllable rEenactment GAN". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (cit. on p. 56).

Umeyama, Shinji (1991). "Least-squares estimation of transformation parameters between two point patterns". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 13.04, pp. 376–380 (cit. on p. 66).

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30 (cit. on pp. 1, 2, 27, 28, 32, 68).

Wang, Chengyi, Chen, Sanyuan, Wu, Yu, Zhang, Ziqiang, Zhou, Long, Liu, Shujie, Chen, Zhuo, Liu, Yanqing, Wang, Huaming, Li, Jinyu, et al. (2023). "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers". In: *arXiv preprint arXiv:2301.02111* (cit. on p. 41).

Wang, Kaisiyuan, Wu, Qianyi, Song, Linsen, Yang, Zhuoqian, Wu, Wayne, Qian, Chen, He, Ran, Qiao, Yu, and Loy, Chen Change (2020). "MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation". In: *ECCV* (cit. on pp. 8, 58, 78).

Wang, Ting-Chun, Liu, Ming-Yu, Zhu, Jun-Yan, Liu, Guilin, Tao, Andrew, Kautz, Jan, and Catanzaro, Bryan (2018a). "Video-to-Video Synthesis". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. (cit. on pp. 14, 73).

Wang, Ting-Chun, Liu, Ming-Yu, Zhu, Jun-Yan, Tao, Andrew, Kautz, Jan, and Catanzaro, Bryan (2018b). "High-resolution image synthesis and semantic manipulation with conditional gans". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807 (cit. on pp. 14, 73).

Wang, Xinsheng, Xie, Qicong, Zhu, Jihua, Xie, Lei, and Scharenborg, Odette (2022). "AnyoneNet: Synchronized Speech and Talking Head Generation for Arbitrary Persons". In: *IEEE Transactions on Multimedia* (cit. on pp. 8, 58).

Wang, Zhou, Bovik, Alan C, Sheikh, Hamid R, and Simoncelli, Eero P (2004). "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4, pp. 600–612 (cit. on p. 61).

Wu, Yuxin and He, Kaiming (2018). "Group normalization". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19 (cit. on p. 56).

Xiao, Zhisheng, Kreis, Karsten, and Vahdat, Arash (2021). "Tackling the generative learning trilemma with denoising diffusion GANs". In: *arXiv preprint arXiv:2112.07804* (cit. on p. 32).

Xing, Jinbo, Xia, Menghan, Zhang, Yuechen, Cun, Xiaodong, Wang, Jue, and Wong, Tien-Tsin (2023). "CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior". In: *arXiv preprint arXiv:2301.02379* (cit. on pp. 7, 55).

Xu, Chao, Zhu, Junwei, Zhang, Jiangning, Han, Yue, Chu, Wenqing, Tai, Ying, Wang, Chengjie, Xie, Zhifeng, and Liu, Yong (2023). "High-fidelity generalized emotional talking face generation with multi-modal emotion space learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6609–6619 (cit. on p. 91).

Yadlin-Segal, Aya and Oppenheim, Yael (2021). "Whose dystopia is it anyway? Deepfakes and social media regulation". In: *Convergence* 27.1, pp. 36–51 (cit. on pp. 21, 91).

Yamagishi, Junichi, Nose, Takashi, Zen, Heiga, Ling, Zhen-Hua, Toda, Tomoki, Tokuda, Keiichi, King, Simon, and Renals, Steve (2009). "Robust speaker-adaptive HMM-based text-to-speech synthesis". In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.6, pp. 1208–1230 (cit. on p. 40).

Yang, Dongchao, Liu, Songxiang, Huang, Rongjie, Lei, Guangzhi, Weng, Chao, Meng, Helen, and Yu, Dong (2023). "InstructTTS: Modelling Expressive TTS in Discrete Latent Space with Natural Language Style Prompt". In: *arXiv preprint arXiv:2301.13662* (cit. on p. 42).

Yang, Ling, Zhang, Zhilong, Song, Yang, Hong, Shenda, Xu, Runsheng, Zhao, Yue, Shao, Yingxia, Zhang, Wentao, Cui, Bin, and Yang, Ming-Hsuan (2022). "Diffusion models: A com-

prehensive survey of methods and applications". In: *arXiv preprint arXiv:2209.00796* (cit. on pp. 3, 32).

Yao, Kaisheng, Zweig, Geoffrey, Hwang, Mei-Yuh, Shi, Yangyang, and Yu, Dong (2013). "Recurrent neural networks for language understanding." In: *Interspeech*, pp. 2524–2528 (cit. on p. 27).

Yao, Xinwei, Fried, Ohad, Fatahalian, Kayvon, and Agrawala, Maneesh (2021). "Iterative text-based editing of talking-heads using neural retargeting". In: *ACM Transactions on Graphics (TOG)* 40.3, pp. 1–14 (cit. on pp. 57, 59).

Ye, Zhenhui, Jiang, Ziyue, Ren, Yi, Liu, Jinglin, Zhang, Chen, Yin, Xiang, Ma, Zejun, and Zhao, Zhou (2023). "Ada-TTA: Towards Adaptive High-Quality Text-to-Talking Avatar Synthesis". In: *arXiv preprint arXiv:2306.03504* (cit. on pp. 8, 58).

Yoshimura, Takayoshi (2002). "Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems". In: *PhD diss, Nagoya Institute of Technology* (cit. on p. 40).

Yosinski, Jason, Clune, Jeff, Bengio, Yoshua, and Lipson, Hod (2014). "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems* 27 (cit. on pp. 16, 78).

Yu, Chengzhu, Lu, Heng, Hu, Na, Yu, Meng, Weng, Chao, Xu, Kun, Liu, Peng, Tuo, Deyi, Kang, Shiyin, Lei, Guangzhi, et al. (2019). "Durian: Duration informed attention network for multimodal synthesis". In: *arXiv preprint arXiv:1909.01700* (cit. on pp. 8, 59).

Yu, Fisher and Koltun, Vladlen (2015). "Multi-scale context aggregation by dilated convolutions". In: *arXiv preprint arXiv:1511.07122* (cit. on p. 40).

Yu, Zhentao, Yin, Zixin, Zhou, Deyu, Wang, Duomin, Wong, Finn, and Wang, Baoyuan (2022). "Talking Head Generation with Probabilistic Audio-to-Visual Diffusion Priors". In: *arXiv preprint arXiv:2212.04248* (cit. on p. 57).

Zhang, Chaoning, Zhang, Chenshuang, Zheng, Sheng, Qiao, Yu, Li, Chenghao, Zhang, Mengchun, Dam, Sumit Kumar, Thwal, Chu Myaet, Tun, Ye Lin, Huy, Le Luang, kim, Donguk, Bae, Sung-Ho, Lee, Lik-Hang, Yang, Yang, Shen, Heng Tao, Kweon, In-So, and Hong, Choong-Seon (2023). "A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need?" In: *arXiv preprint arXiv:2303.11717* (cit. on pp. 24, 64).

Zhang, Kaipeng, Zhang, Zhanpeng, Li, Zhifeng, and Qiao, Yu (2016). "Joint face detection and alignment using multitask cascaded convolutional networks". In: *IEEE signal processing letters* 23.10, pp. 1499–1503 (cit. on pp. 9, 65).

Zhang, Mingyuan, Cai, Zhongang, Pan, Liang, Hong, Fangzhou, Guo, Xinying, Yang, Lei, and Liu, Ziwei (2022a). "Motiondiffuse: Text-driven human motion generation with diffusion model". In: *arXiv preprint arXiv:2208.15001* (cit. on p. 32).

Zhang, Sibo, Yuan, Jiahong, Liao, Miao, and Zhang, Liangjun (2022b). "Text2video: Text-Driven Talking-Head Video Synthesis with Personalized Phoneme-Pose Dictionary". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2659–2663 (cit. on p. 58).

Zhang, Tao (2022). "Deepfake generation and detection, a survey". In: *Multimedia Tools and Applications* 81.5, pp. 6259–6276 (cit. on p. 91).

Zhang, Wenxuan, Cun, Xiaodong, Wang, Xuan, Zhang, Yong, Shen, Xi, Guo, Yu, Shan, Ying, and Wang, Fei (2022c). "SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation". In: *arXiv preprint arXiv:2211.12194* (cit. on pp. 8, 17, 56, 81).

Zhang, Zhimeng, Li, Lincheng, Ding, Yu, and Fan, Changjie (2021). "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670 (cit. on pp. 16, 78, 79).

Zhou, Yi, Wu, Chenglei, Li, Zimo, Cao, Chen, Ye, Yuting, Saragih, Jason, Li, Hao, and Sheikh, Yaser (2020). "Fully convolutional mesh autoencoder using efficient spatially varying kernels". In: *Advances in neural information processing systems* 33, pp. 9251–9262 (cit. on p. 50).

Zielonka, Wojciech, Bolkart, Timo, and Thies, Justus (2022). "Towards metrical reconstruction of human faces". In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*. Springer, pp. 250–269 (cit. on pp. 6, 50).

Zollhöfer, Michael, Thies, Justus, Garrido, Pablo, Bradley, Derek, Beeler, Thabo, Pérez, Patrick, Stamminger, Marc, Nießner, Matthias, and Theobalt, Christian (2018). "State of the art on monocular 3D face reconstruction, tracking, and applications". In: *Computer graphics forum*. Vol. 37. 2. Wiley Online Library, pp. 523–550 (cit. on p. 48).