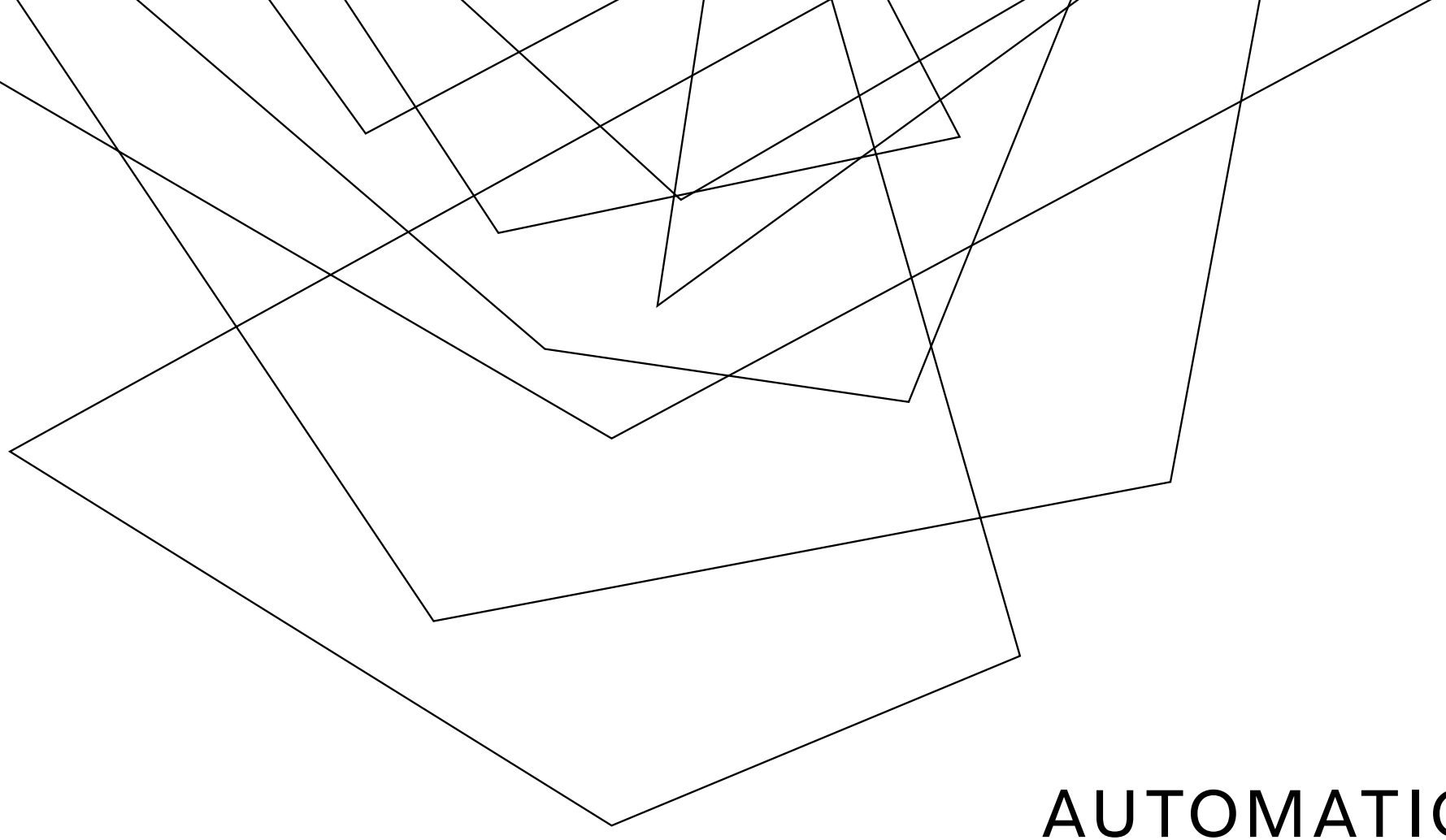


AUTOMATIC SPEECH RECOGNITION

Μάριος Σαρρηγιάννης

ΕΙΣΑΓΩΓΗ

Σκοπός είναι να αναπτυχθεί ένα μοντέλο Speech to Text (STT) το οποίο θα λαμβάνει ως είσοδο ένα σήμα ήχου και θα το μετατρέπει σε κείμενο.

DATASET

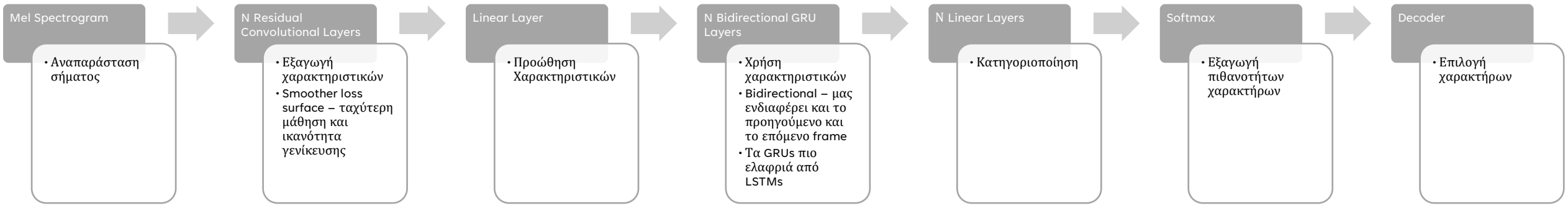
- Χρησιμοποιήθηκε το LibriSpeech ASR Corpus και συγκεκριμένα το clean-100 version (χωρίς θόρυβο) το οποίο περιλαμβάνει δεδομένα εκπαίδευσης 100 ωρών ομιλίας μαζί με την απομαγνητοφώνηση τους
- Περιλαμβάνει 125 γυναίκες και 126 άντρες ομιλητές, 25 λεπτά για τον καθένα
- Η χρήση αυτού του dataset μας επιτρέπει να συγκρίνουμε τα αποτελέσματα μας με άλλα ASR μοντέλα

ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ

- Η ακρίβεια των ASR μοντέλων μετριέται με 2 μετρικές:
  - Word Error Rate (WER)
  - Character Error Rate (CER)
- Παίρνουν τιμές στο [0, 1] με το 0 να είναι το βέλτιστο
- WER (ή CER) = (S + D + I) / N = (S + D + I) / (S + D + C) όπου
  - S: εναλλαγές
  - D: διαγραφές
  - I: εισαγωγές
  - C: σωστά
  - N: συνολικός αριθμός (λέξεων ή χαρακτήρων)

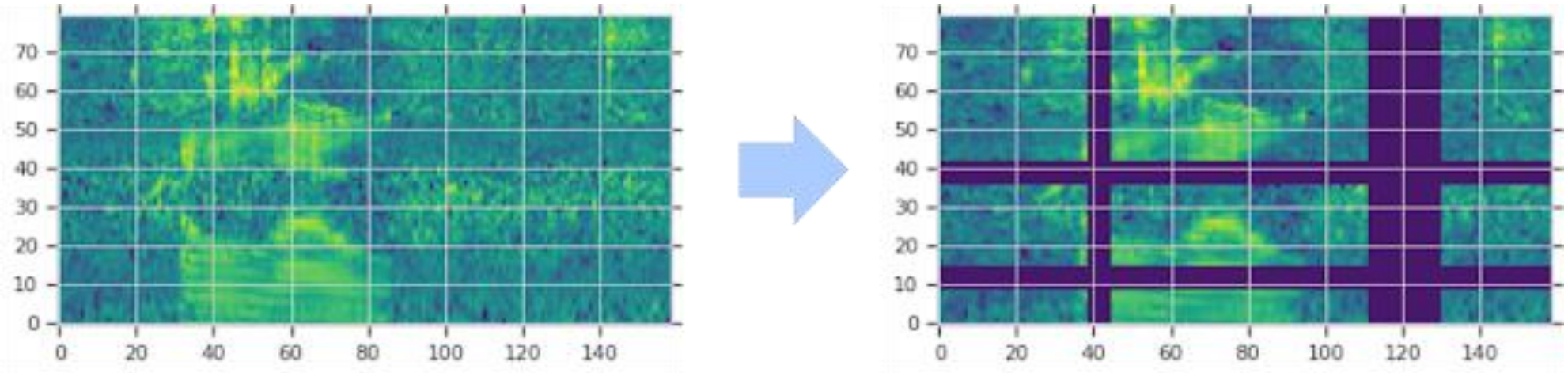
ΑΡΧΙΤΕΚΤΟΝΙΚΗ

Η αρχιτεκτονική βασίζεται στο Deep Speech 2



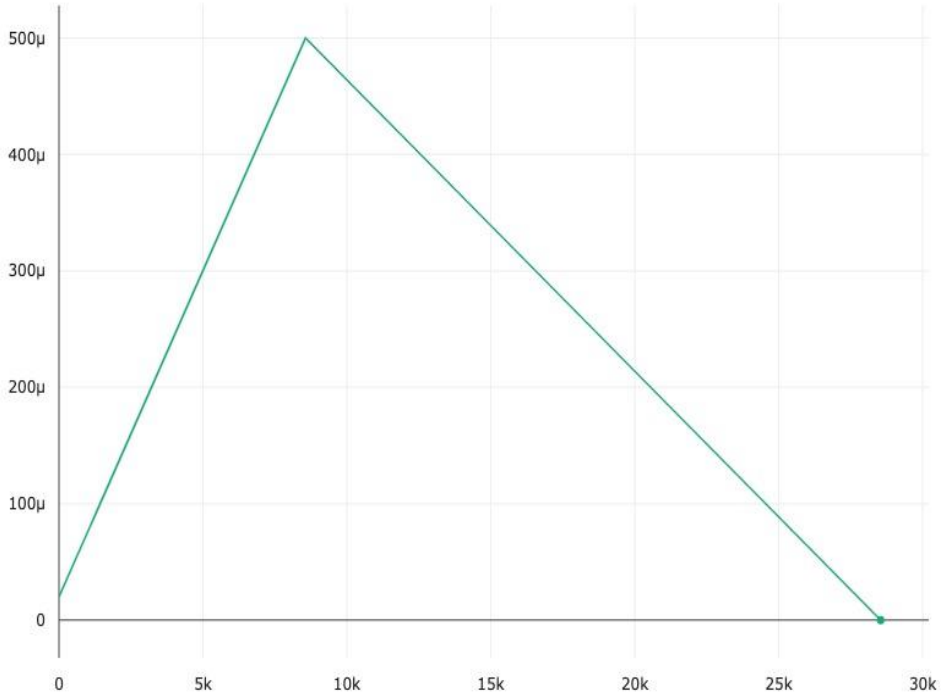
DATA AUGMENTATION

- Χρησιμοποιείται η τεχνική Spectrogram Augmentation (SpecAugment)
- Αφαιρούνται ομοιόμορφα συχνότητες και χρονικά διαστήματα
- Βελτιώνει τη γενίκευση και αποτρέπει το overfitting



OPTIMIZER - SCHEDULER

- Χρησιμοποιείται η βελτιωμένη έκδοση του Adam, ο AdamW, ο οποίος χρησιμοποιεί weight decay αντί για L2 normalization
- Χρησιμοποιείται ο One Cycle Learning Rate Scheduler, οποίος αυξάνει δραστικά το learning rate και μετέπειτα το μειώνει, βελτιώνοντας την ταχύτητα και την ικανότητα γενίκευσης του μοντέλου



DECODING-TEMPORAL ALIGNMENT

- Χρησιμοποιείται ένας greedy decoder ο οποίος διαλέγει τον χαρακτήρα με την μεγαλύτερη πιθανότητα σε κάθε frame και στη συνέχεια ενώνει τους επαναλαμβανόμενους χαρακτήρες και αγνοεί τον «κενό» χαρακτήρα
- Για την αποφυγή του temporal alignment γίνεται χρήση του Connectionist Temporal Classification (CTC) Loss Function

ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

- Το μοντέλο εκπαιδεύτηκε για 5 εποχές με 1 ResCNN και 1 BiGRU layer (λόγω limitations του colab), με χρόνο εκπαίδευσης 50 λεπτών
- WER: 0.6978
- CER: 0.2674

- Παρατηρούμε ότι προβλέπει σχετικά καλά τους χαρακτήρες αλλά όχι τόσο καλά τις λέξεις

ΠΑΡΑΔΕΙΓΜΑΤΑ ΕΚΤΕΛΕΣΗΣ

Model: the hoped they would be sto heordiner turnips an caret send brsed betat hos and that butoan peaces to bed elatle doutin the thick pepered flower fan sos

Actual: he hoped there would be stew for dinner turnips and carrots and bruised potatoes and fat mutton pieces to be ladled out in thick peppered flour fattened sauce

Model: stuffid in to you his belly countul him

Actual: stuff it into you his belly counselled him

ΣΥΜΠΕΡΑΣΜΑΤΑ – ΕΠΕΚΤΑΣΕΙΣ

- Το μοντέλο παράγει ενθαρρυντικά αποτελέσματα για το μέγεθός του (4.760.733 παράμετροι) και για τις λίγες εποχές εκπαίδευσης
- Για το πρόβλημα σύνθεσης λέξεων, μια πιθανή λύση θα ήταν να χρησιμοποιηθούν labels λέξεων/συλλαβών με τα μειονεκτήματα ότι θα αύξανε δραματικά τις απαιτήσεις του μοντέλου σε μνήμη και θα δημιουργούσε ισχυρή εξάρτηση στην ποιότητα του vocabulary
- Μια ακόμα λύση στο παραπάνω πρόβλημα θα ήταν η χρήση NLP τεχνικών – π.χ. το κείμενο που παράγεται να διορθώνεται με τη βοήθεια ενός προ-εκπαιδευμένου transformer
- Το ASR πρόβλημα πλέον προσεγγίζεται καλύτερα με transformers (wav2vec 2.0)

ΕΥΧΑΡΙΣΤΩ!

Ερωτήσεις;