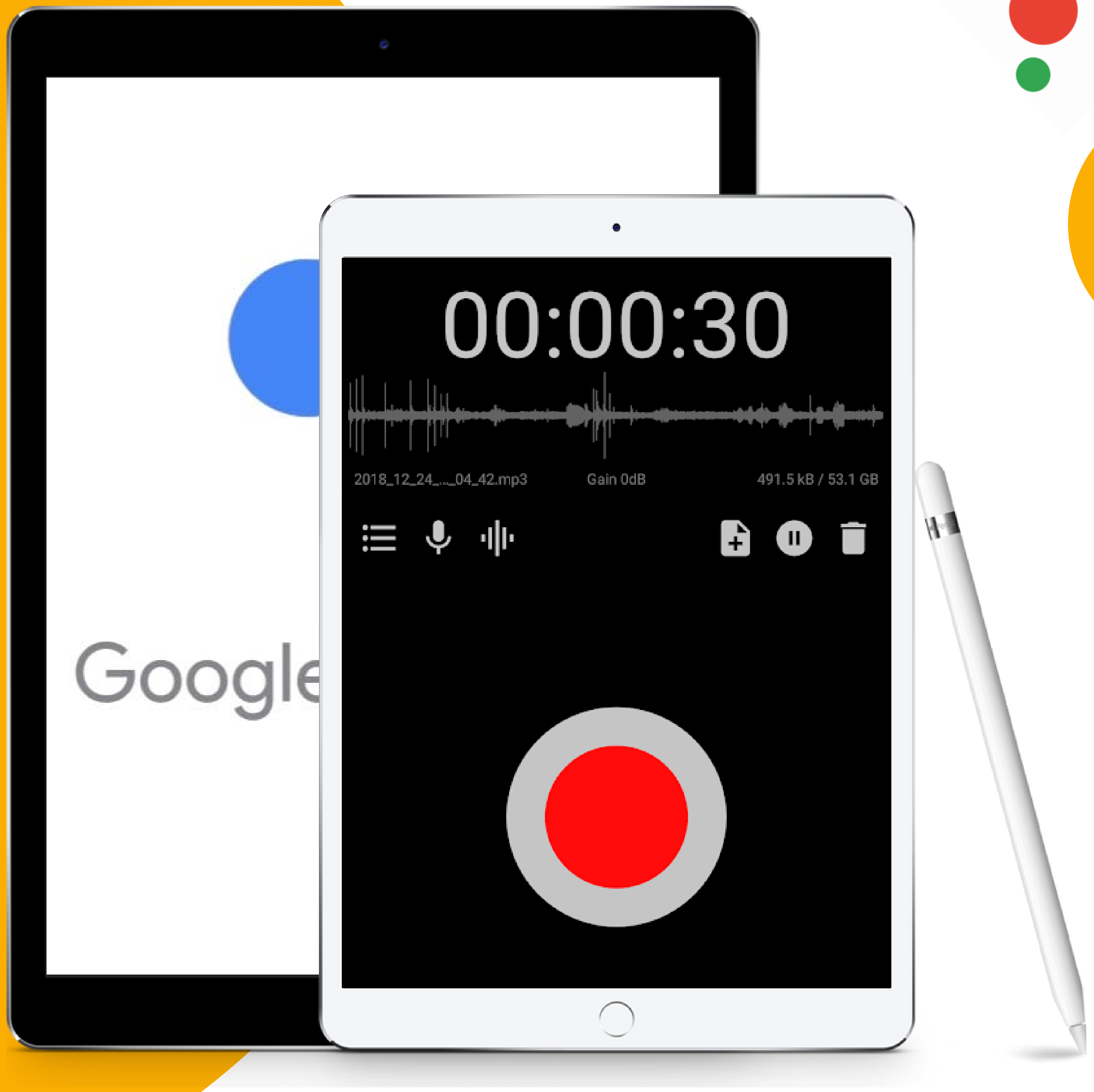
Get started with Speech Recognition using TensorFlow

Nourchene Ferchichi

Research Engineer

ML Google Developer Expert

Introduction



How does information travel from our brains into the computer?

Introduction

Automatec Speech Recognition, Defined

Computer Science
+
Linguistics

Transforms the spoken word into the written one!

In this Presentation

Here's what we'll cover:

1. Applications
2. SR Dataset
3. DNN for Acoustic modeling
4. Language Models
5. Practice

Applications of ASR

Telecome Industry

Conversational Chatbots to
Enhance customer support
Resolve Technical Issues
Provide Personalised Advice

Marketing

Voice Search for accurate
product search

Internet of Things

Smart TVs
Autonumus cars
Hands-free help at home

Applications of ASR

Health Care

Digital Assistant:

Medical guidance
Quick access to
administrative information

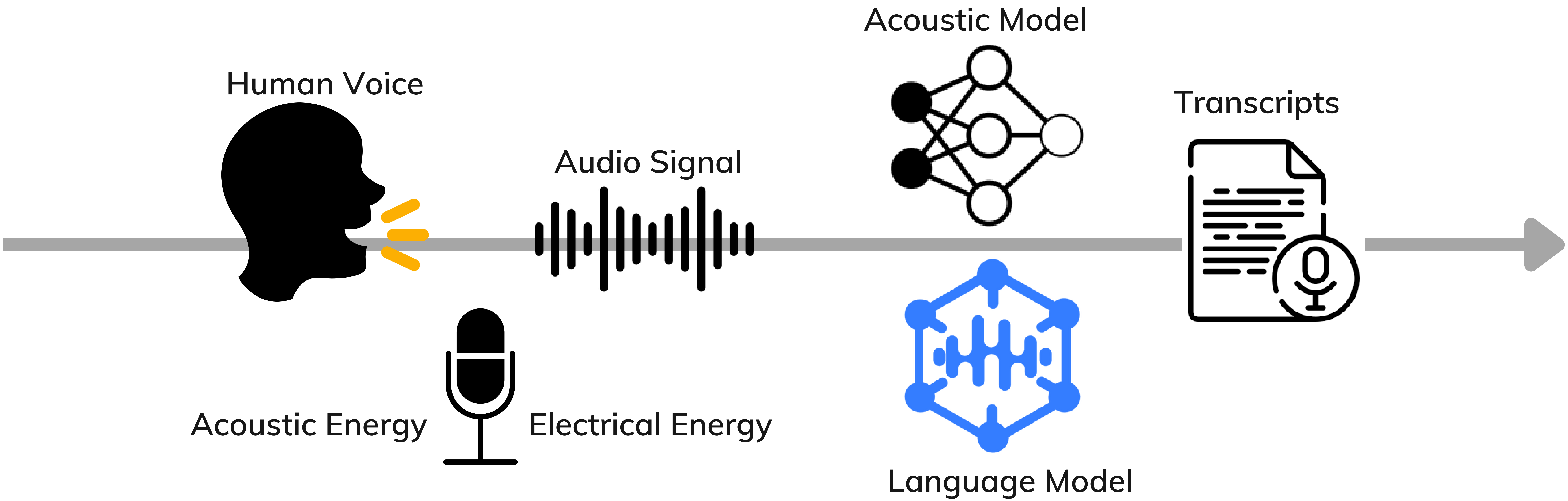
Banking

Request transaction
information
Make payments

Workplace

Transcribe real-time
conferences
Report meetings

Pipeline



Human Voice
Audio Signal
Acoustic Model
Transcripts
Acoustic Energy
Electrical Energy
Language Model

Datasets format

Features (X)	Labels (y)
	*Good Morning!*
Audio wave	Transcript

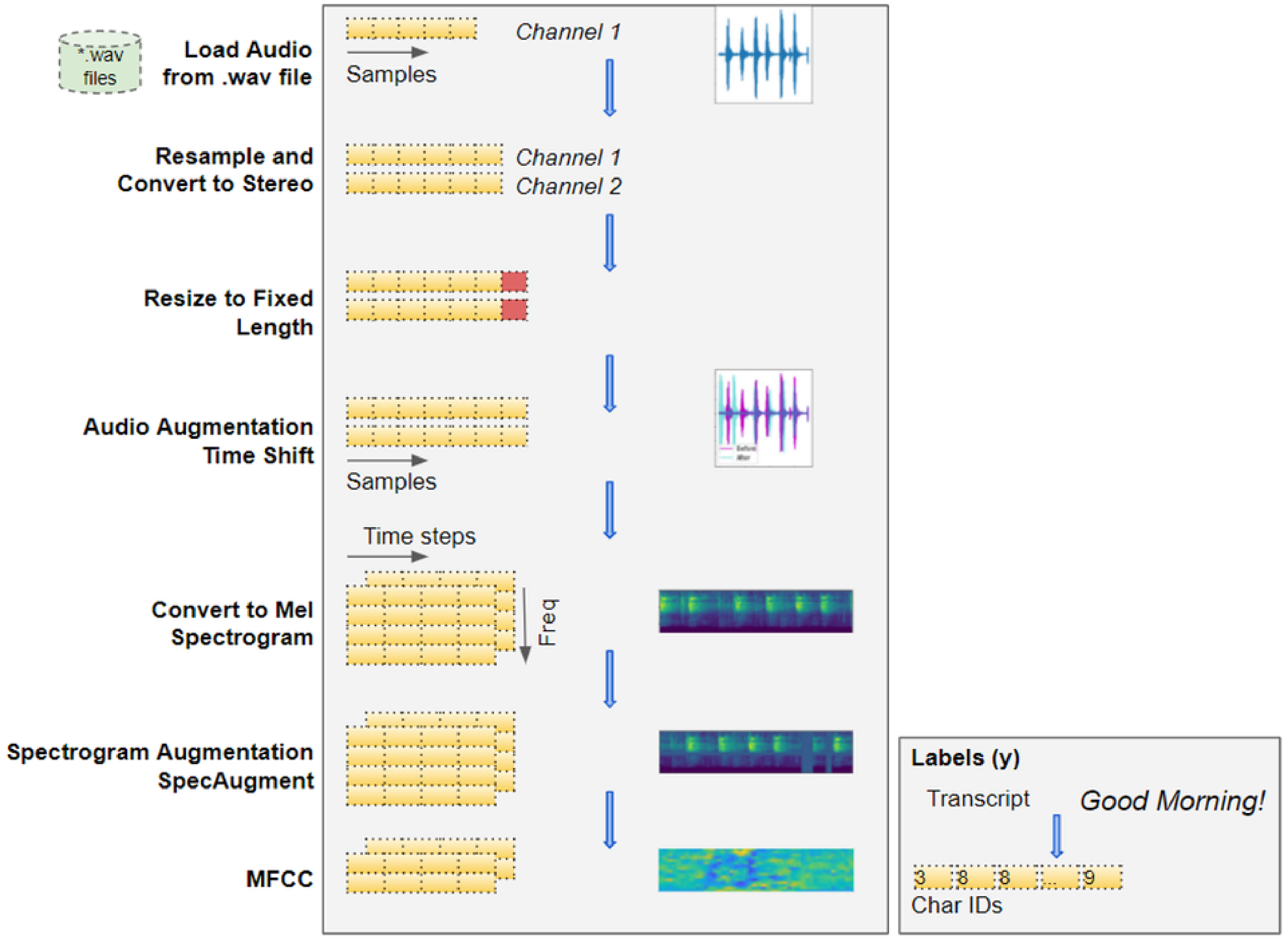
Audio Waves: Features

Audio clips of spoken sequences

Transcripts: Labels

Text transcripts of what was spoken

Datasets Preprocessing:
Audio

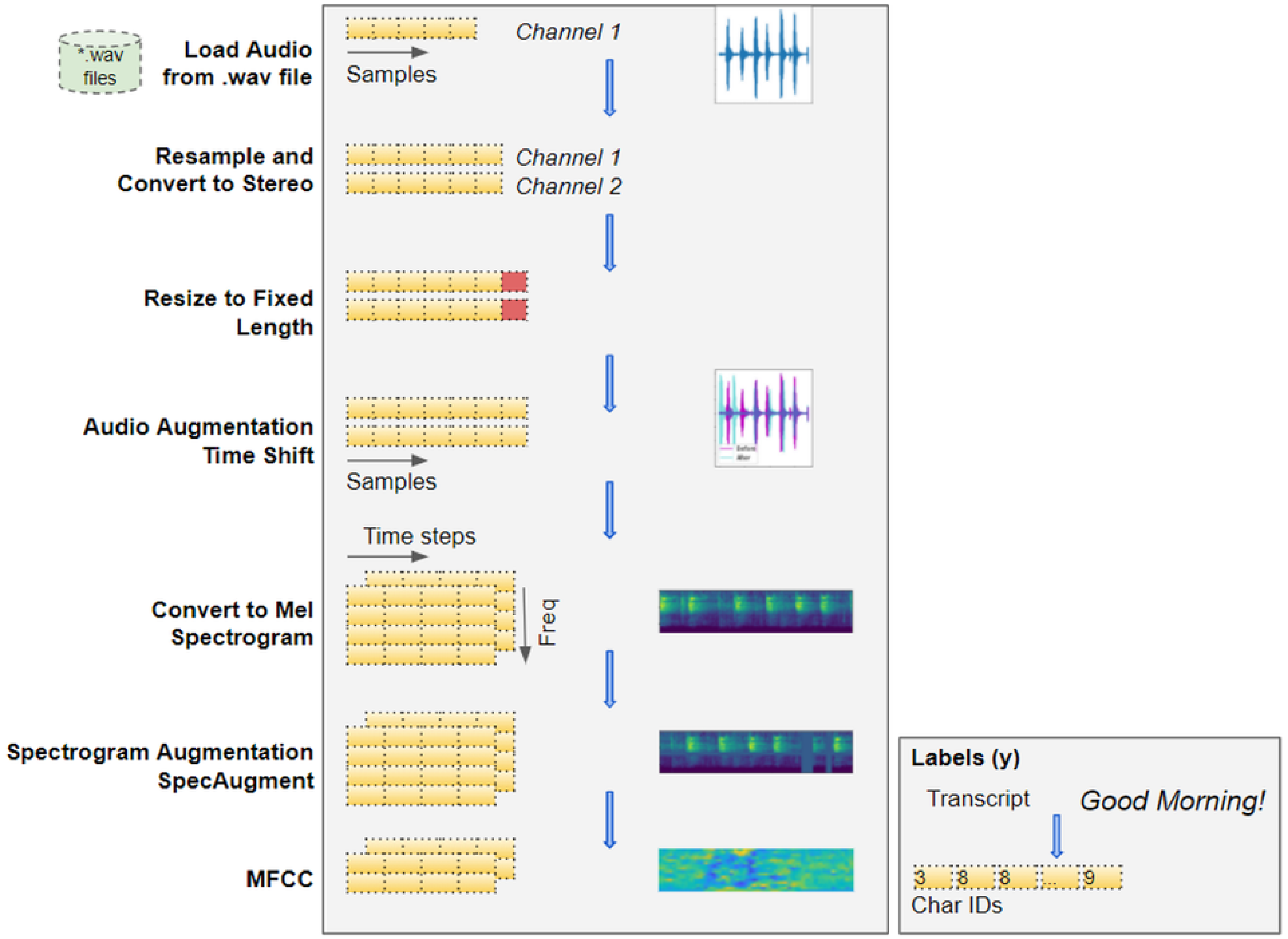


Datasets Preprocessing:

Audio

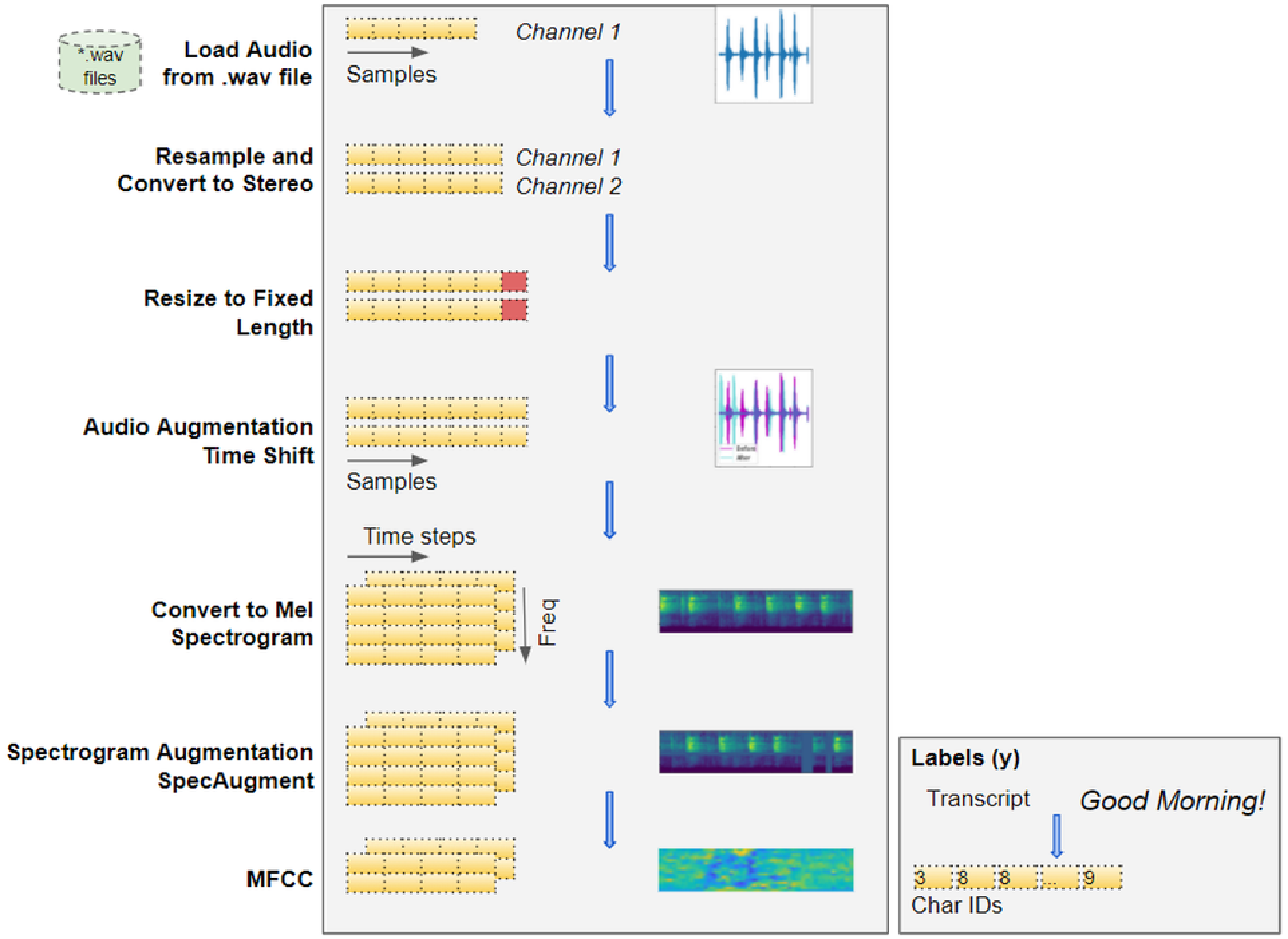
1.Convert to uniform dimensions

Standardize the dimensions of our audio data



Datasets Preprocessing:

Audio



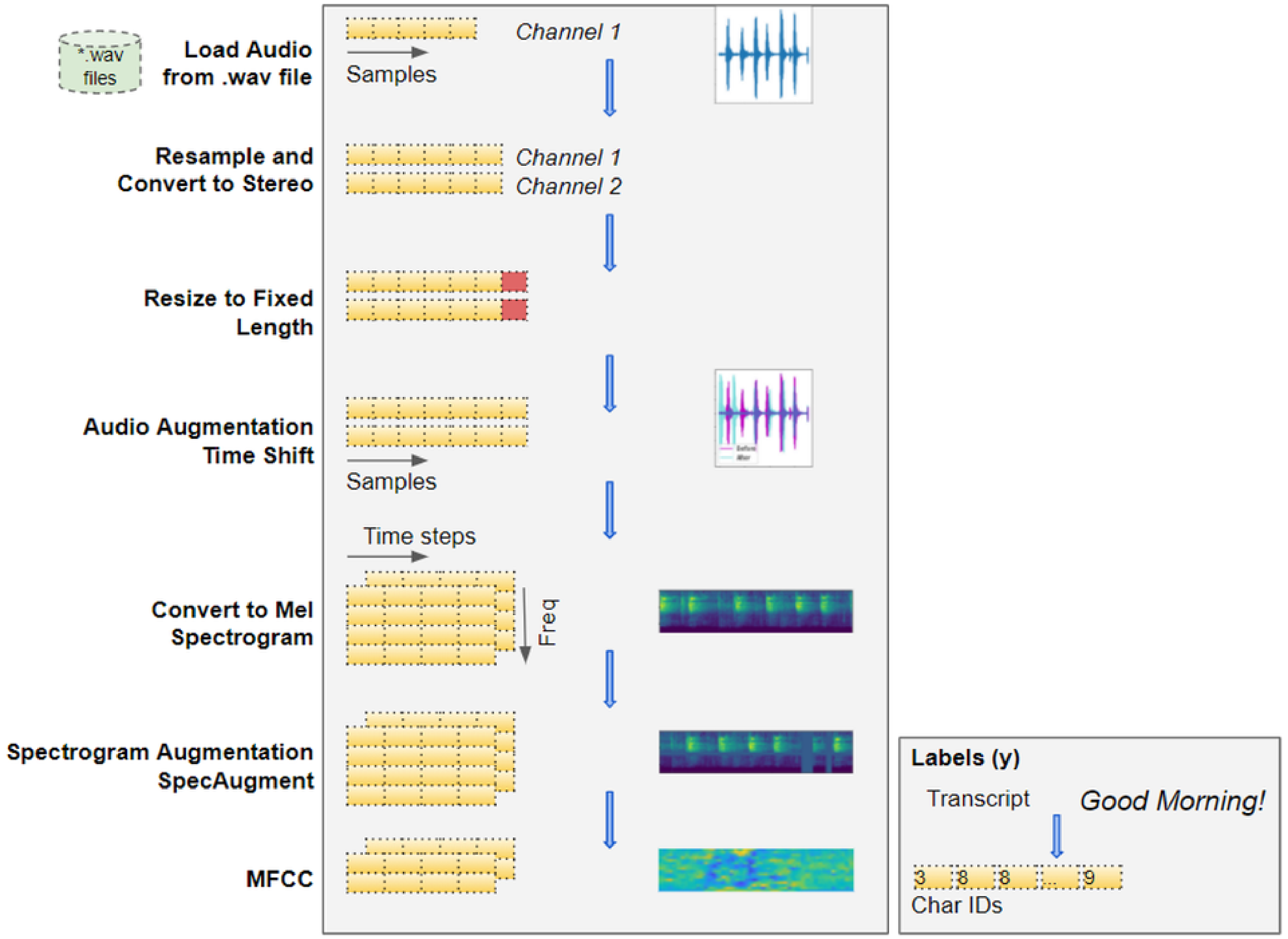
1.Convert to uniform dimensions

Standardize the dimensions of our audio data

2. Data Augmentation of raw audio

Add more variety to our input data

Datasets Preprocessing:
Audio



1.Convert to uniform dimensions

Standardize the dimensions of our audio data

2. Data Augmentation of raw audio

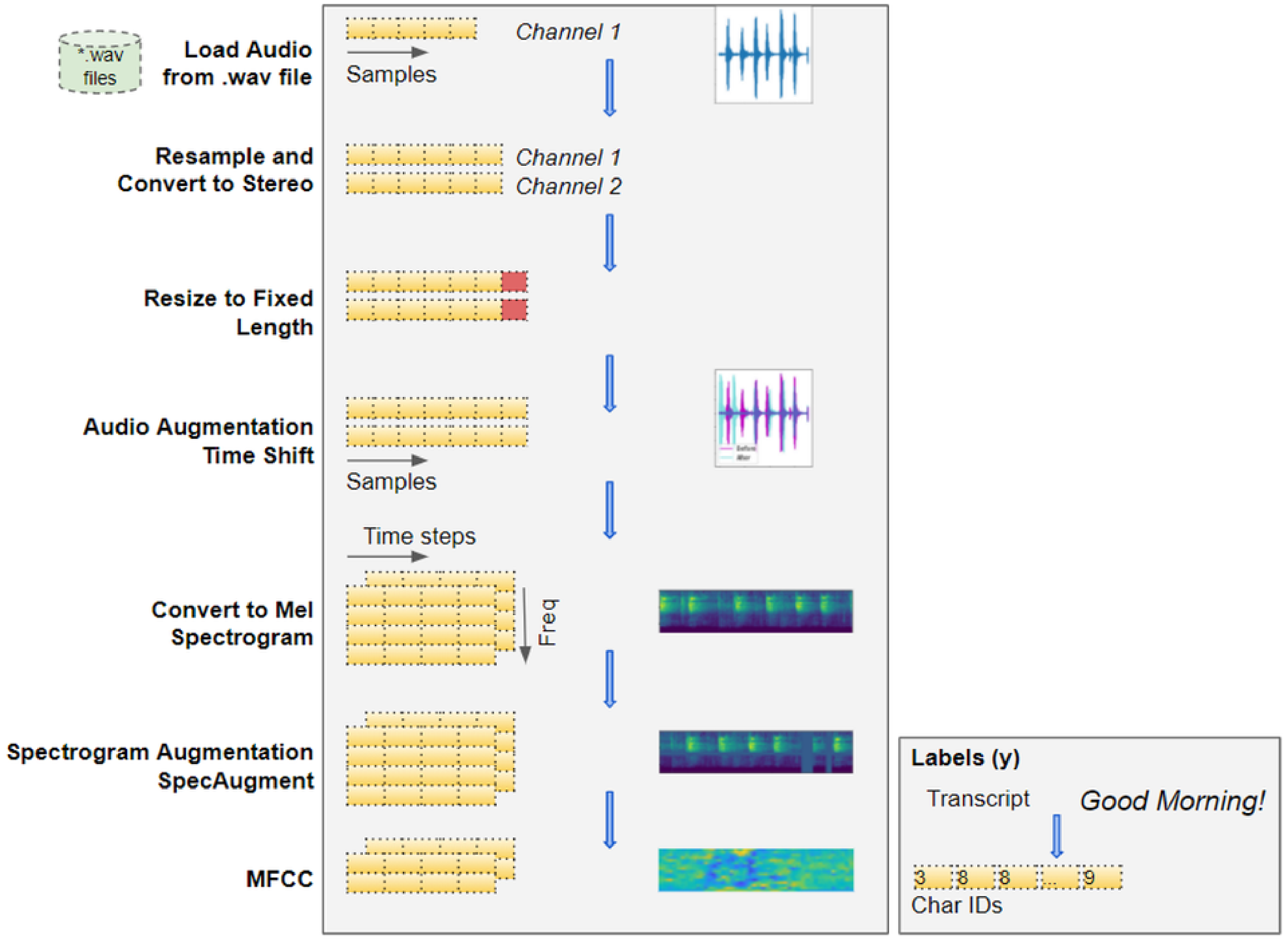
Add more variety to our input data

3. Mel Spectrograms

Captures the nature of the audio as an image

Datasets Preprocessing:

Audio



1.Convert to uniform dimensions

Standardize the dimensions of our audio data

2. Data Augmentation of raw audio

Add more variety to our input data

3. Mel Spectrograms

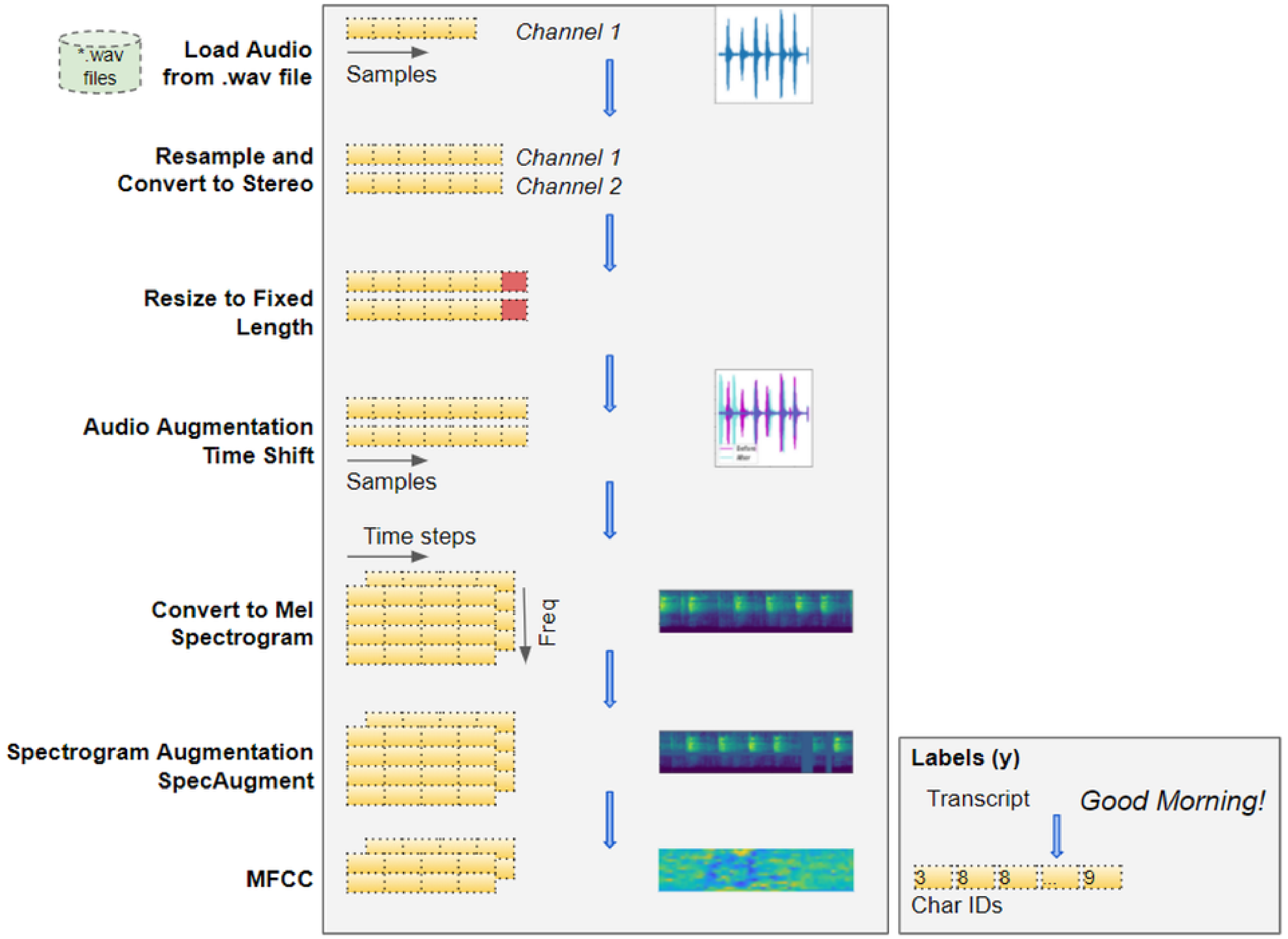
Captures the nature of the audio as an image

4. MFCC

Extract the most essential frequency coefficients

Datasets Preprocessing:

Audio



1.Convert to uniform dimensions

Standardize the dimensions of our audio data

2. Data Augmentation of raw audio

Add more variety to our input data

3. Mel Spectrograms

Captures the nature of the audio as an image

4. MFCC

Extract the most essential frequency coefficients

5. Data Augmentation of Spectrograms

Apply random Frequency and Time Masking

Datasets Preprocessing: Transcripts

Normalization

Bring the words to their closer to a predefined "standard"

Punctuation Removal

Remove punctuation characters like !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~ from a text

Buid the vocabulary

Build a vocabulary from each character in the transcript and convert them into character IDs.

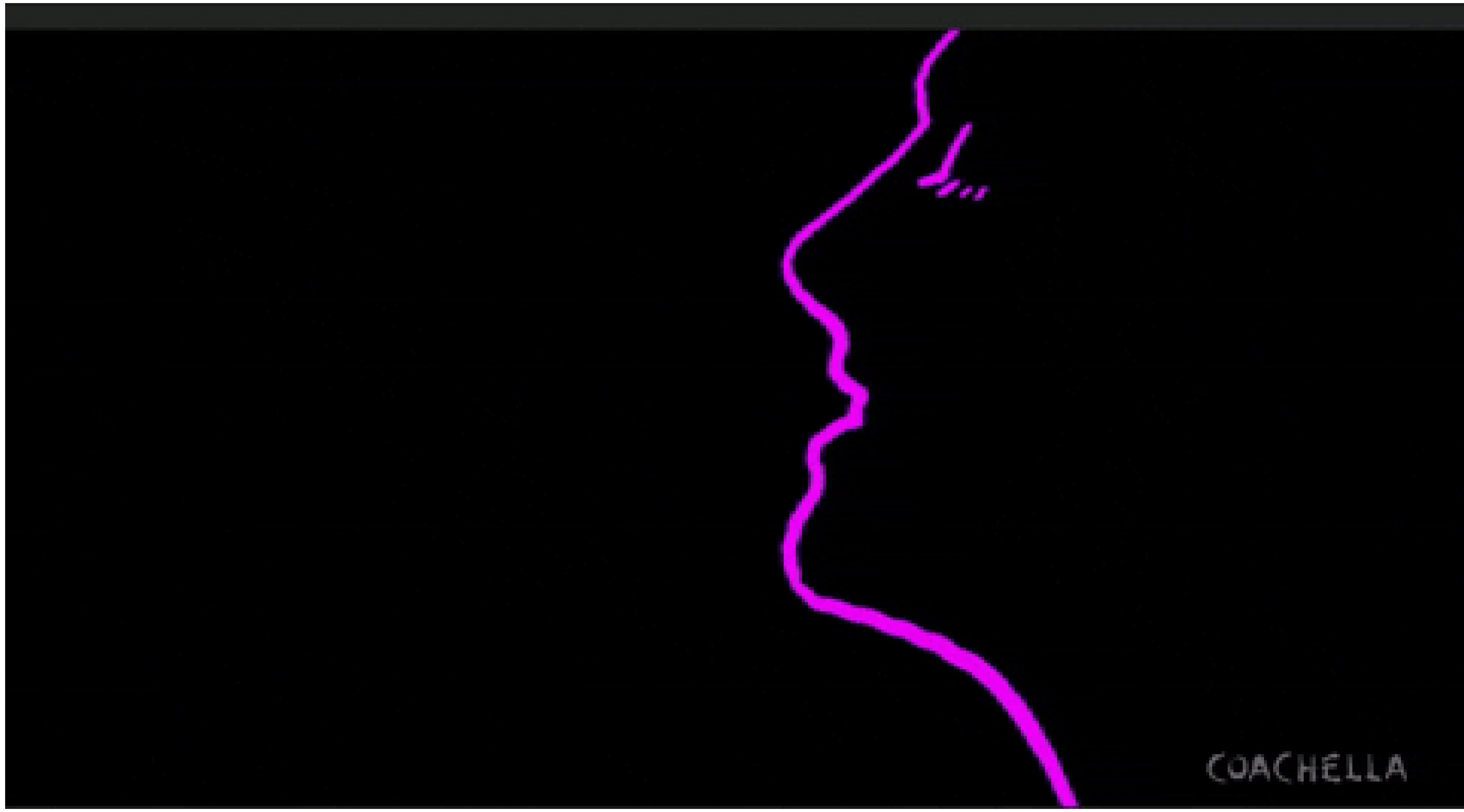
Datasets exaples

AudioMNIST

Libriispeech

LJ Speech

VoxForge

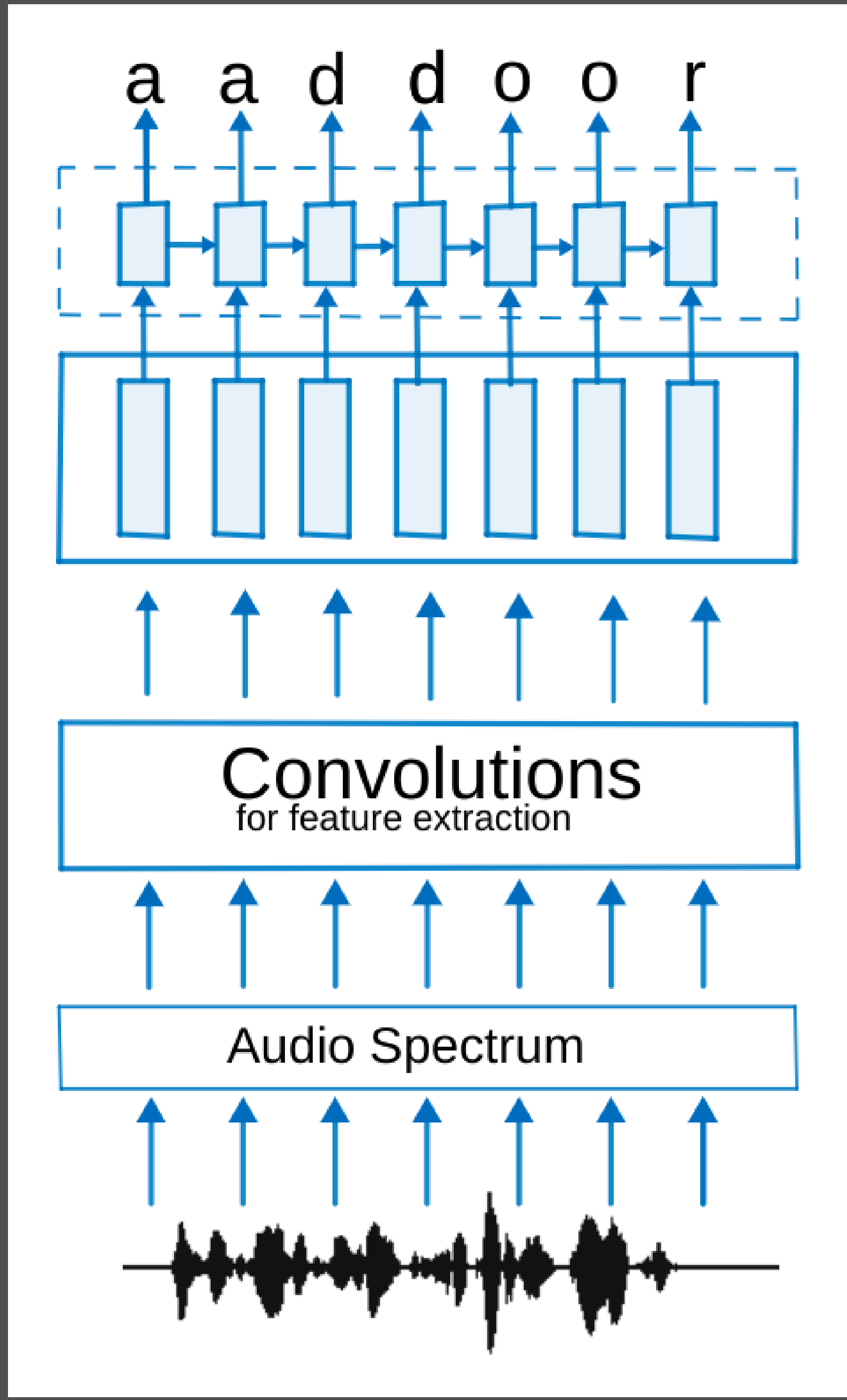


DNN for Acoustic modeling

Different Types

1. A CNN plus RNN-based architecture that uses the CTC
2. RNN-based sequence-to-sequence network

DNN for Acoustic modeling
Model architecture Overview



4 Connectionist Temporial Classification Decoder

3 A Linear Layer

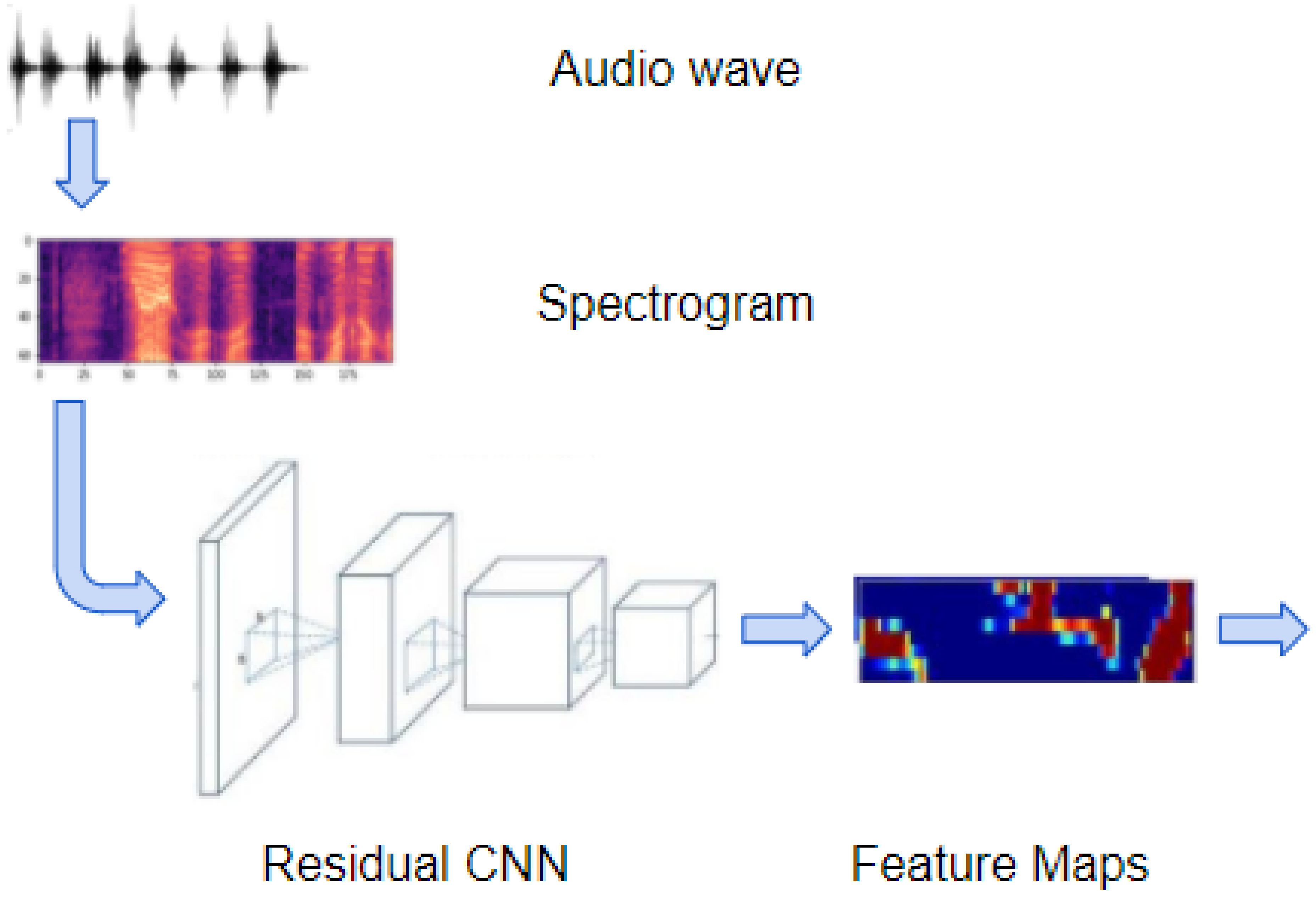
2 Recurrent Neural Network

1 Convolutional Neural Network

DNN for Acoustic modeling

Model architecture Explained

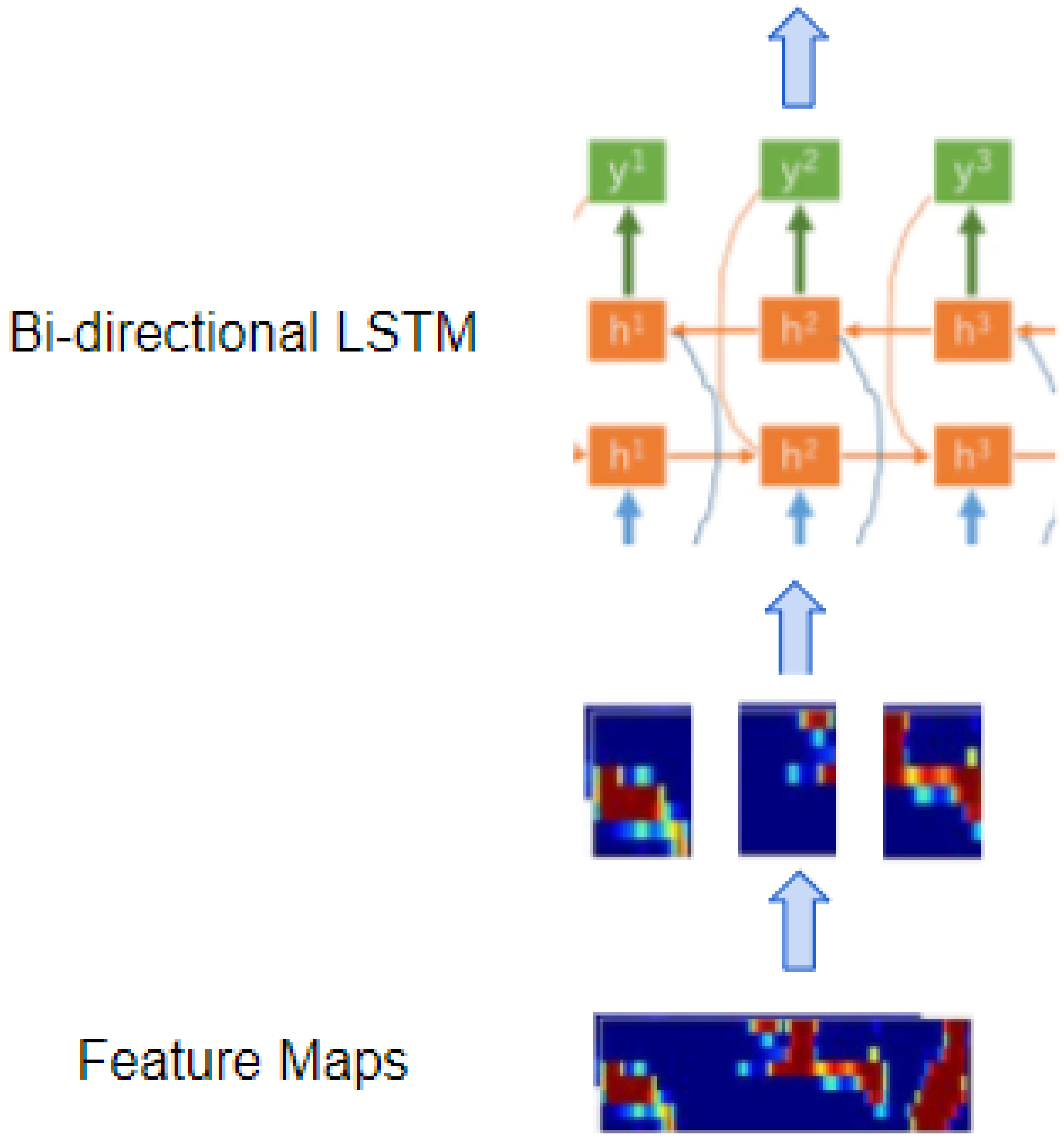
1 Convolutional Neural Network



DNN for Acoustic modeling

Model architecture Explained

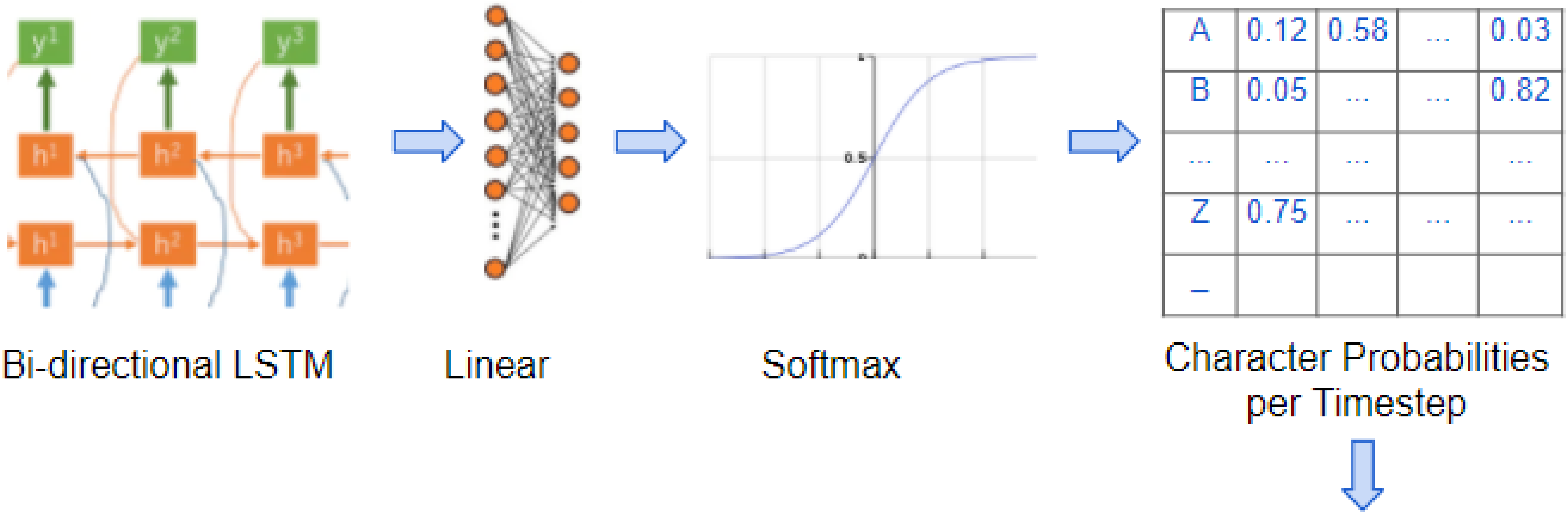
2 Recurrent Neural Network



DNN for Acoustic modeling

Model architecture Explained

3 A Linear Layer

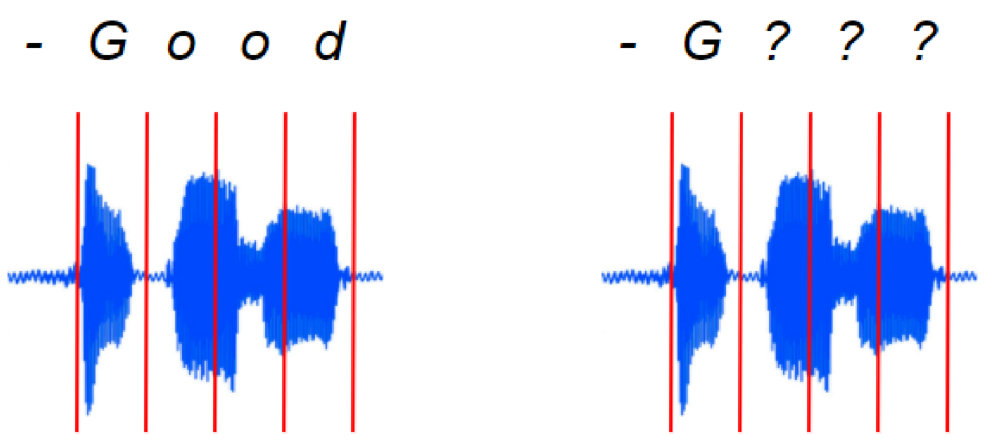


A	0.12	0.58	...	0.03
B	0.05	...	...	0.82
...	...	...		...
Z	0.75	...	...	...
_				

DNN for Acoustic modeling

Model architecture Explained

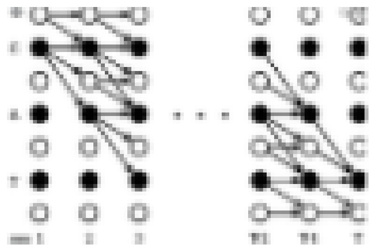
4 Connectionist Temporial Classification



A	0.12	0.58	...	0.03
B	0.05	...	...	0.82
...	...	...		...
Z	0.75	...	...	...
–				

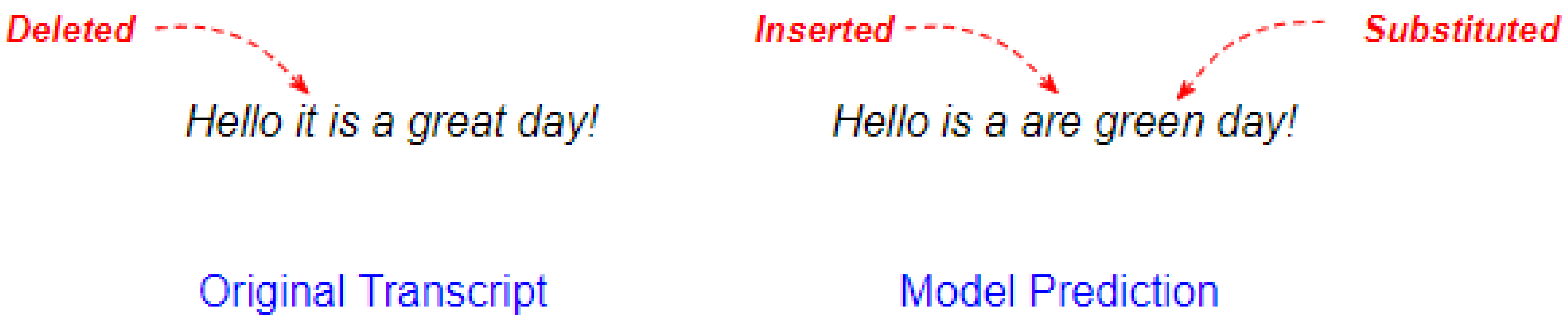
Character Probabilities per Timestep

Decoding



Output

Good Morning!



$$\text{Word Error Rate} = \frac{\text{Inserted} + \text{Deleted} + \text{Substituted}}{\text{Total words in transcript}}$$

$$= \frac{1 + 1 + 1}{6}$$

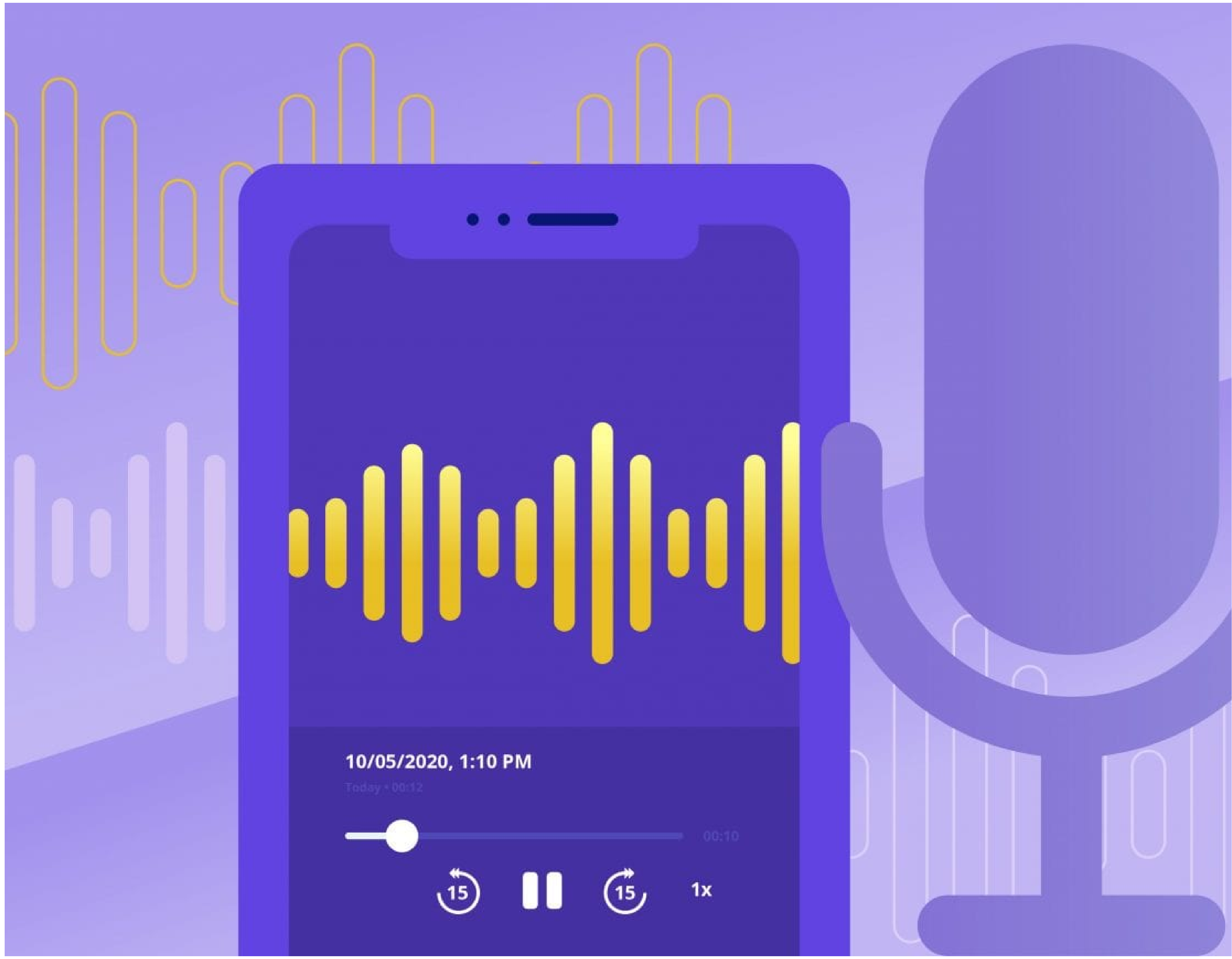
$$= 0.5$$

Metrics — Word Error Rate (WER)

The metric formula is fairly straightforward. It is the percent of differences relative to the total number of words.

Adding a Language Model

Great speech to text AI requires a great language model we recognize what words we predict, as well as pronunciation models to handle differences between accents, dialects, age, gender, and the many other factors that make our voices unique.

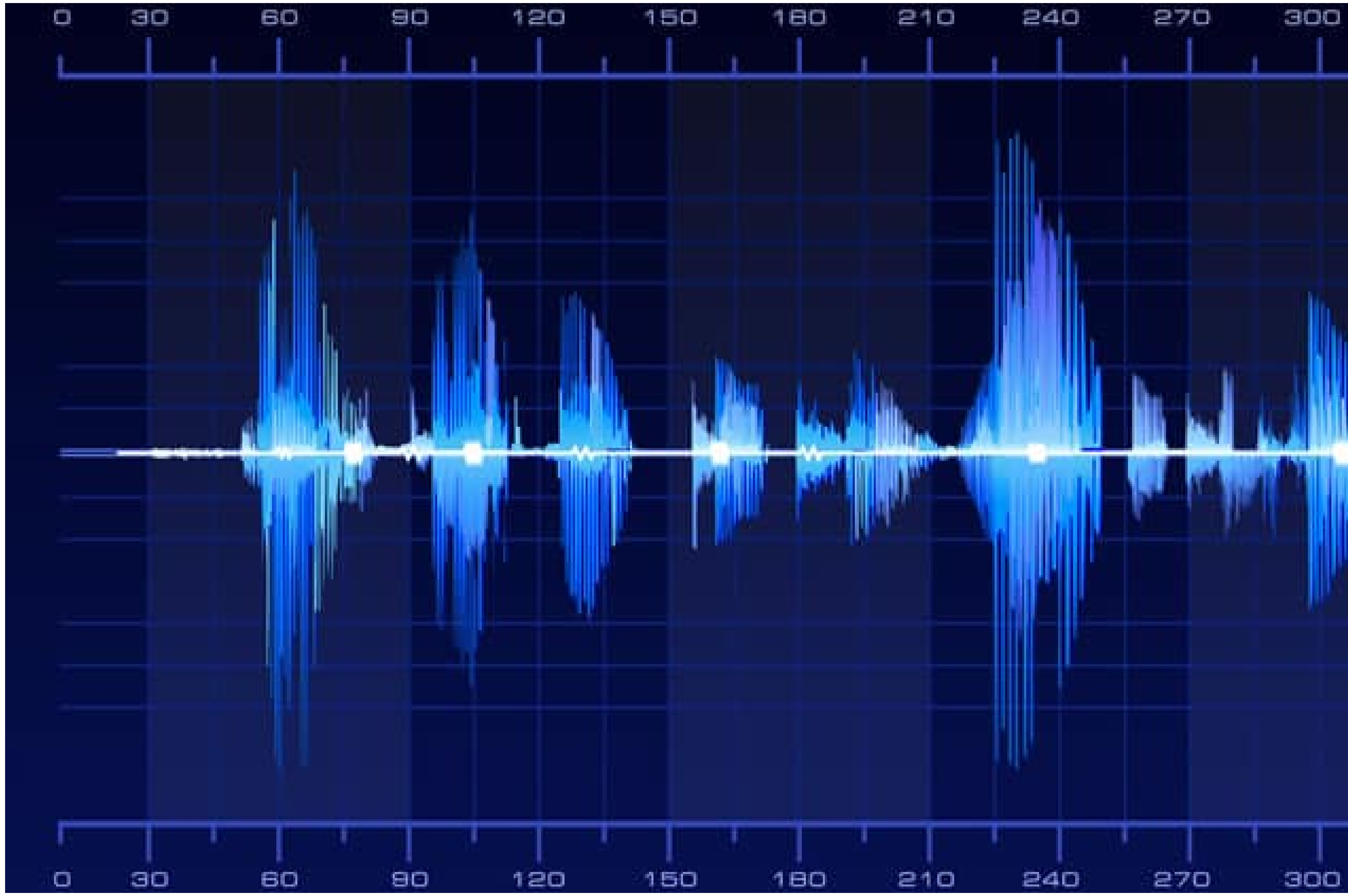


Inverse Normalization

Reverse Normalization

Adding puntuation

Adding capitalization



"The advance of technology is based on making it fit in so that you don't really even notice it, so it's part of everyday life.”

Bill Gates

Practice Time!

Do you have any questions?

We hope you learned something new.