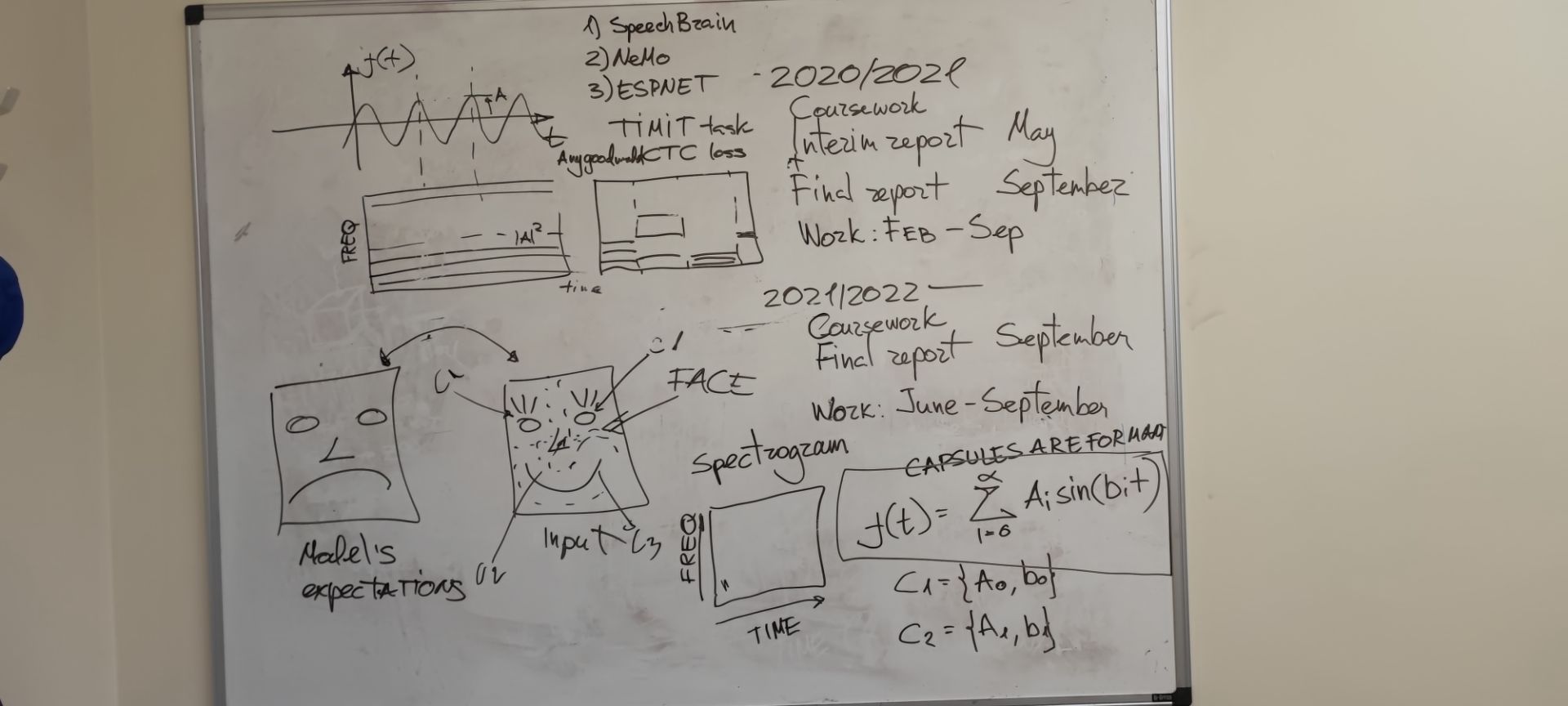
**Overview:**

****

**Spectrogram图像解析**

<https://home.cc.umanitoba.ca/~krussll/phonetics/acoustic/spectrogram-sounds.html>

Timeline

Description automatically generated

**音频文件可视化为Spectrogram图像**

网站：<https://www.dcode.fr/spectral-analysis>

软件：spectrogram音频可视化.zip，运行exe，将音频拖入软件即可。

**英语发音的spectrogram和formant（\*）**

<https://corpus.eduhk.hk/english_pronunciation/index.php/our-praat-manual/>

**The relation between spectrogram and phoneme**

<https://www.youtube.com/watch?v=qKFx1-xbE-c>

**Speech Process课程**

<https://speech.zone/courses/speech-processing/>

**Digital-Formant Synthesizer for Speech-Synthesis Studies论文笔记**

本文介绍了一个terminal-analog synthesizer，大概意思是模拟真正发生器官的合成器。

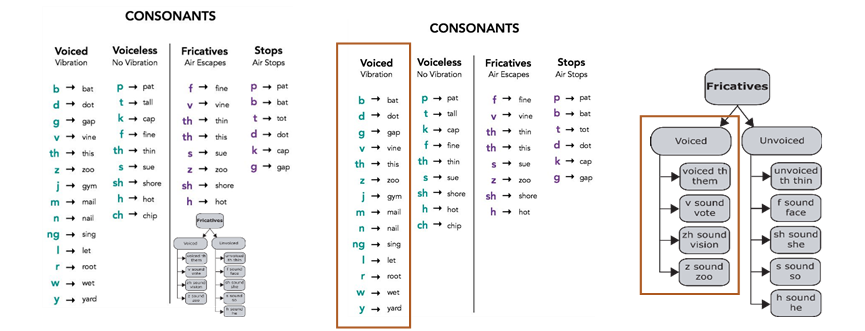
Diagram, schematic

Description automatically generated

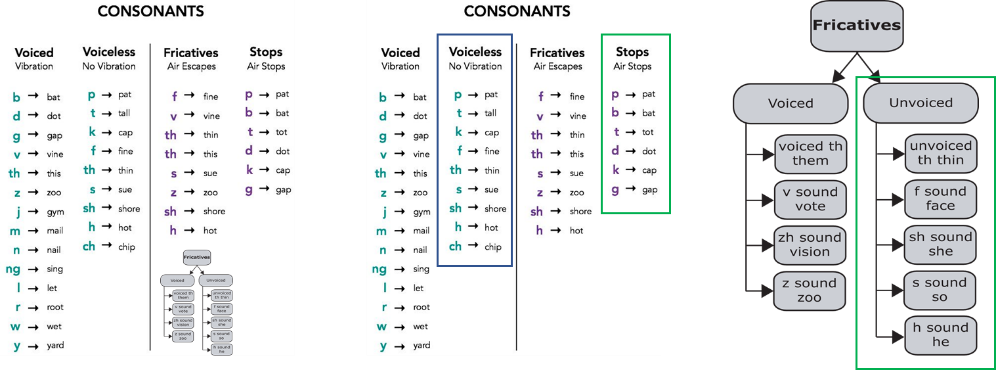
有两个激励源：pitch-impulse generator（音高脉冲发生器）和frication generator（摩擦发生器）。

Pitch-impulse generator由外部信号F0指定的速率发生脉冲。其可以生成**浊音（voiced**

**speech）**和**浊擦音(voiced fricatives)**的**浊音(voiced)和清音(unvoiced)成分**的激励。



Frication generator生成**吐气声音（aspiration）**、**耳语（whispered speech）**、**擦音（fricative）**和**清塞音（voiceless stop consonant）**的**摩擦噪声（frication noise）**。



其中，绿色为确定的Frication generator生成目标，蓝色不确定。

**Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis论文笔记**

**SPEECH-TRANSFORMER: A NO-RECURRENCE SEQUENCE-TO-SEQUENCE MODEL FOR SPEECH RECOGNITION 论文笔记**

As a no-recurrence seq2seq model, the Speech-Transformer differs from recurrent seq2seq models mainly on two aspects:

Firstly, both the encoder and decoder are composed of multi-head attention and position-wise feed-forward networks rather than RNNs.

第一，全部使用多头注意力和按位置的前馈网络，而不是RNN

Secondly, the encoder outputs h are attended by each decoder block respectively, replacing the one-step intermediary attention of recurrent seq2seq models.

第二，encoder的输出h会被分别加到每个解码器上，而不是循环seq2seq的one-step中间注意力

Model Structure:

Input: Spectrogram

CNN for feature extraction

Additional modules to extract more expressive representations

2D-Encoder

Positional Encoding

Decoder

[**总览：李宏毅深度 s学习人类语言处理课程**](https://www.youtube.com/playlist?list=PLJV_el3uVTsO07RpBYFsXg-bN5Lu0nhdG)

[**ASR survey**](https://www.researchgate.net/profile/Imran-Makhdoom/publication/345710977_Automatic_speech_recognition_a_survey/links/5fbc7321299bf104cf6e85c0/Automatic-speech-recognition-a-survey.pdf)

**History**

* Since the early 1950s, researchers have been trying to design ASR.
* In the early 1980s, the hidden Markov model (HMM) was introduced.
* In the later years of the 1980s, the n-gram model was introduced.
* In the early years of the 2000s, the HMM was being used in combination with a feed-forward artificial neural network (ANN).
* Nowadays, long-short term memory (LSTM), a type of recurrent neural network (RNN), is being used for speech recognition in combination with different deep learning techniques.

**Evaluation Metrics**

Speed

Real-time factor (RTF) = the time of processing (P)/duration of the input audio (I)

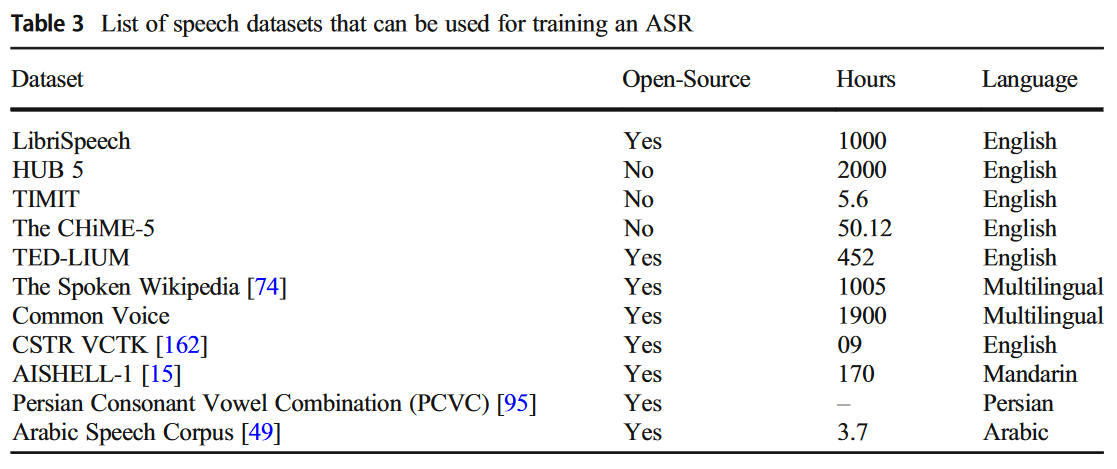
Accuracy

Word error rate (WER) = (substituted word num (S) + missed word num (D) + extra word num (I))/ ground truth word num (N)

Word recognition rate (WRR) = (correctly predicted word num (H) – I)/N = 1-WER

Character error rate (CER)

**Datasets**



**Model Architecture**

**Pre-processing module** (reduce the signal-to-noise ratio)

e.g., framing, normalization, end-point detection and pre-emphasis

**Feature extraction module**

e.g., Mel-frequency cepstral coefficients (MFCCs), linear predictive coding (LPC), and discrete

wavelet transform (DWT)

**Classification model**

Legacy Models: GMM, SVM and ANN

New Models:

1. [Listen, Attend and Spell (LAS)](https://www.youtube.com/watch?v=BdUeBa6NbXA)

Listen is kind of a encoder including LSTM, RNN, CNN and Self-Attention. 用于特征提取

由于序列很长，因此还会采用Down-Sampling技术，来减少数据长度，RNN有Pyramid-RNN和Pooling over time方法，CNN有Time-delay DNN方法，Self-attention有Truncated Self-attention。

Attend is Attention.

Spell is kind of a decoder.

\*Beam Search

2. [CTC](https://www.youtube.com/watch?v=L519dCHUCog)

CTC是一个输入一个输出，每个输入之间是独立的，Recurrent Neural Aligner (RNA)考虑了输入之间的关系，将前一个的输出作为后一个的输入。

同时，CTC无法解决单输入多输出的情况，比如th。

一种解决方法是直接把th当作一个结果，一个class，简单粗暴。

还有通过网络的设计解决的方法是RNN-T，其一个输入可以有多个输出，直到网络认为输出完成为止。

3. Transducers

和RNN-T类似，只不过这次是多输入多输出，每次输入多个信号，然后全部预测结束再预测下一组。

4. MoChA

相比前面的模型，每次输入的Input是一个一个或者一组一组输入的，MoChA使用了类似滑动窗口的机制，让网络自己判断窗口是否往前移动，并且使用了特别的技巧让网络可以训练。

5. Transformer

6. Wav2sec

[**Language model**](https://www.youtube.com/watch?v=dymfkWtVUdo&list=PLJV_el3uVTsO07RpBYFsXg-bN5Lu0nhdG&index=8) (form trigrams, words or sentences)

LM is to estimate the probability of token sequence.

ASR部分设计

使用Transformer，目测不用Set Transformer，因为语音是有顺序的，并且顺序不能忽略

台大李宏毅21年机器学习课程 self-attention和transformer：

<https://www.bilibili.com/video/BV1Xp4y1b7ih>

Transformer一般会考虑用NAT做并行计算，但是性能有下降，并且导致multi-modality

**TTS survey**

Formant synthesis is a kind of source-filter model.

WaveNet is a popular source-filter model.