# CS753
# Keyword Spotting for Literacy Assessment

Shreeharsha B S (18307R002)
Sachin Nayak (17307R002)

## TASK DEFINITION

Given an input speech utterance and a keyword, the output is a yes/no decision as to whether the keyword/s of interest was spotted or not. Multiple keywords and repeated keywords are also spotted if they occur in the test speech sample provided.

## METHODOLOGY

An Attention-based Recurrent Sequence Generator(ARSG), which predicts a phone sequence for a given utterance is used.[1] This seq2seq model consists of 4 bidirectional GRUs for the Encoder and 1 GRU unit for the Decoder.

Different input feature methodologies were used in the model :
i) **The spectrogram** of the signal, this is processed by a CNN (size:5,8,32,2) which generates an embedding for the signal which is then sent to the Encoder which is supposed to help learn attention weights more easily

ii) **39 MFCC** features for each 20 ms window and 10 ms frame of the signal. These features by-pass the CNN and go directly to the encoder

iii) **39 Wavelet MFCC** features for each 10 ms frame. These features also by pass the CNN.
They were calculated by doing wavelet decomposition on the given speech signal at 2 stages to get CA2, CD2, CD1 (II stage approximation and detail coefficients, I stage detail coefficients). These signals are sub-sampled such that CD1 is half the original signal length and CA2 and CD2 are one-fourth the original signal length, because they cannot contain any more information than the signal itself, so an un-subsampled signal is redundant.
This is better in theory than the conventional MFCC calculation because when using a fixed window size, frequencies which have time periods less than the window length are leaked into other “bins” while finding the FFT. On the other hand, increasing the window length to avoid leakage leads to poor time resolution and doesn’t work well for non-stationary signals. So, the wavelet decomposition which can be roughly thought of as finding MFCCs on the low pass filtered and high pass filtered versions of the signal might have less errors.

Then, the MFCCs of these CA1, CD2, CA2 are calculated as usual, the coefficients are made back to the original signal length by zero padding. But no delta or double delta features are calculated. This means we have a fair comparison of the features used, however the advantage of the wavelet decomposition to decompose further more stages hasn’t been fully used. The number of stages which of decomposition that is relevant and useful depends on the signal complexity.

The encoder-decoder system learns the mapping between input feature sequence and the output phone sequence label. It also learns the length of the output label that it must produce as output because the loss is CEL(Cross entropy) so the sequence lengths (of the decoder output and the label) must match. While decoding it produces a sequence of phones upon which beam search (with beam size = 10) is done to get the optimal output sequence length. The attention parameters can be learnt and used in many ways: such as localized to some context, making it independent of the previous decoder hidden state, or previous attention weight or hybrid attention as described in [1]. In this work the hybrid attention model was used.

Given the output from the decoder i.e. a sequence of phones, the given keyword is loaded up and all possible pronunciations are found from TIMIT's 39 phone Lexicon by converting the phone sequence which was part of the 60 phone set (which has details ,such as different types of closures that are not relevant to the task of keyword spotting).

## SPOTTING THE KEYWORD

Then these pronunciations are scanned for in the decoded sequence of phones to search for and declare whether a keyword was found. A measure of similarity between two phone sequences similar to Edit Distance is used to make the decision of whether the given keyword was spotted or not. The edit distance wasn't used directly because it is a hard classification and does not take into account the similarity between two different phones. It is not a fine measure, so the edit distance is modified such that phones in similar groups have a cost equaling the inverse of the size of the group if the decoded phone and the expected phone are in the same group. Otherwise, the substitution cost is '1' as before. Then we fix a threshold on the cost and say that a keyword is spotted if the decoded phone sequence is within some percentage of the actual phone sequence. This factor is made adaptive by scaling it with the length of the sequence of phones in the keyword.

The threshold fixed ($ED_{thr}$) is also a measure of how confident the system is when it has spotted a keyword. Higher the factor, the less confident the system is about the keyword and vice-versa.

**To summarize**:
Edit Distance (ED) between phones in the same group:

$ED(a,\tilde{a})= 1/(\text{size of group})$

Otherwise:

$ED(a,b) = 1$

If '$k_1,k_2,k_3,...k_n$' is one possible phone sequence of the keyword, then  we say a sequence of phones '$l_1,l_2,l_3...l_m$' contains the keyword, if $ED((l_i,l_{i+1},...l_{i+n},)(k_1,k_2,k_3,...k_n)) <= 0.3*(n)$

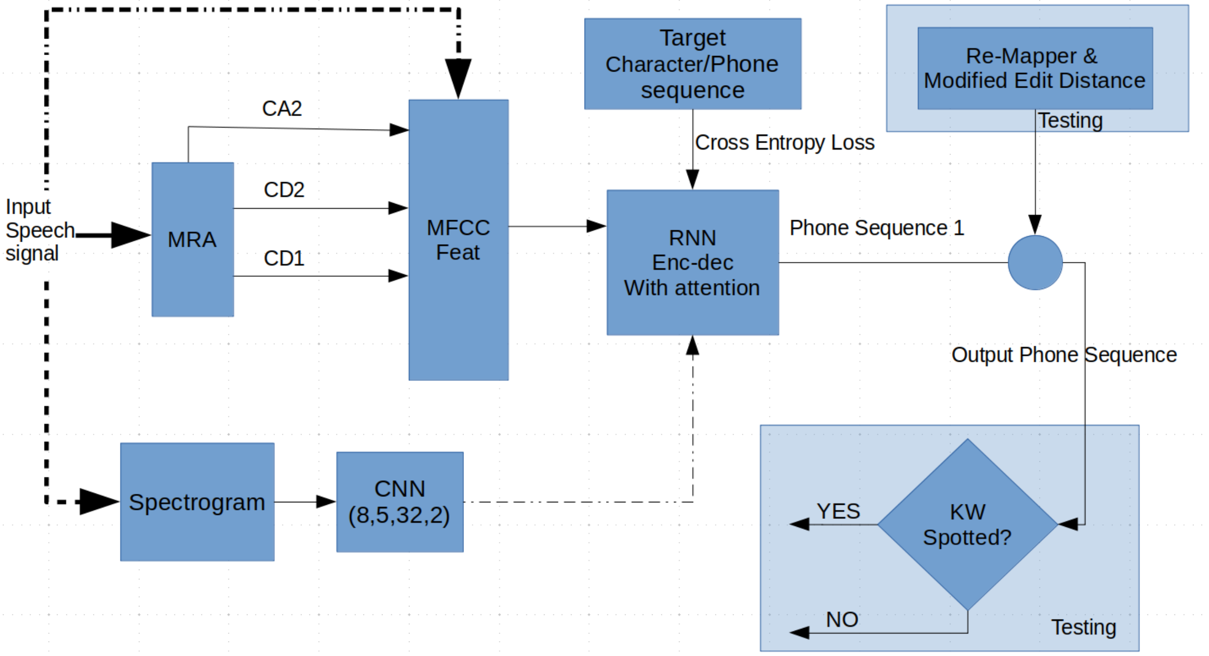


Figure 1.

## DATA-SET

TIMIT is used to train the model. This is an apt database because it has frame level phone markings, which are easily learnt by the CEL function. The details are as follows:

Training set - 3696 recordings over 436 speakers
Development set - 400 recordings over 50 speakers
Test set - 944 recordings over 140 speakers

## TESTING SCENARIO

13 keywords were selected so as to have a good variation in their frequency of occurrence in the data-set. Figure 2. shows the details.

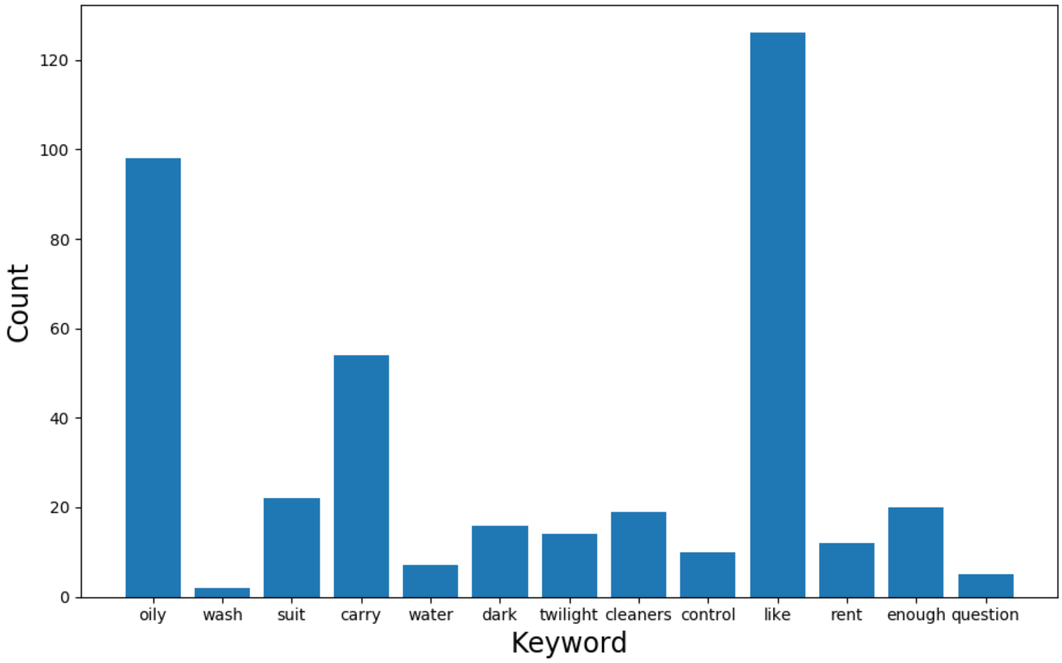


Figure 2.

Overall, the 13 keywords had 148 different pronunciations in total, in the 39 phone space.

## EVALUATION MEASURE

The TPR $= \dfrac{\text{(\# of times the keyword was spotted)}}{\text{\# of times the keyword was in an utterance}}$

and FPR $= \dfrac{\text{(\# of times the keyword was spotted)}}{\text{\# of times the keyword was not in an utterance}}$

is found and averaged over all pronunciations over all keywords.

# EXPERIMENTS

1) Comparing the 3 types of feature extractions on clean TIMIT test set

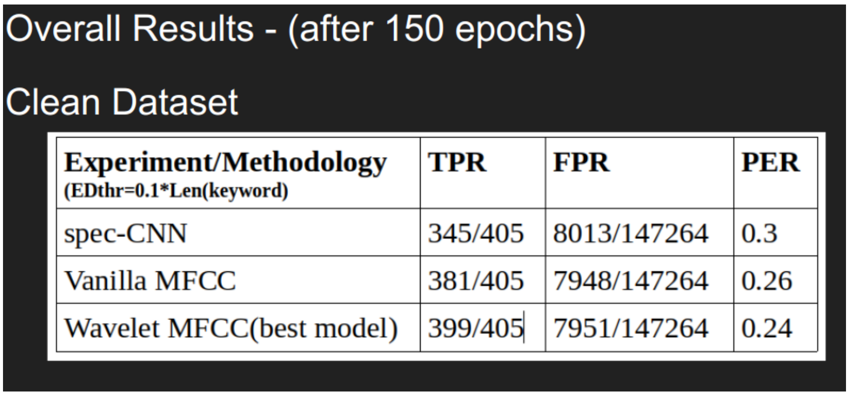
**Using the best configuration from above:**

* Evaluate the model on a degraded data-set (babble noise at 10 dB SNR was artificially added to the training and test set)

* Evaluate the model by combining the noisy and clean data sets (Multi- condition training)

* Evaluate the model on a completely different data-set - Recordings from the Campus school here

# RESULTS AND DISCUSSION



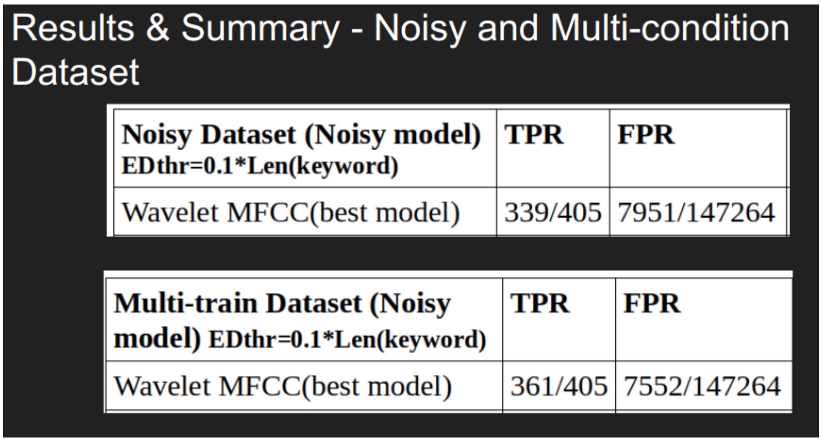
Overall Results - (after 150 epochs)

Clean Dataset

| **Experiment/Methodology** (EDthr=0.1*Len(keyword) | **TPR** | **FPR** | **PER** |
|---|---|---|---|
| spec-CNN | 345/405 | 8013/147264 | 0.3 |
| Vanilla MFCC | 381/405 | 7948/147264 | 0.26 |
| Wavelet MFCC(best model) | 399/405 | 7951/147264 | 0.24 |

The Wavelet based MFCCs seem to have better Phone error rate(PER) over all 39 phones as well as in addition to better TPR and FPR. This suggests the improvement was not just due to the keywords selected.

Further examination is required.

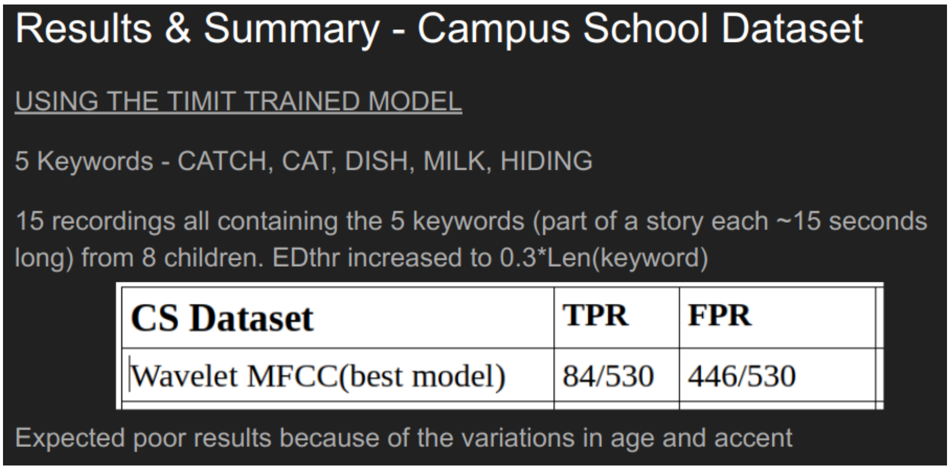


Results & Summary - Noisy and Multi-condition Dataset

| **Noisy Dataset (Noisy model)** **EDthr=0.1*Len(keyword)** | **TPR** | **FPR** |
|---|---|---|
| Wavelet MFCC(best model) | 339/405 | 7951/147264 |

| **Multi-train Dataset (Noisy model)** **EDthr=0.1*Len(keyword)** | **TPR** | **FPR** |
|---|---|---|
| Wavelet MFCC(best model) | 361/405 | 7552/147264 |

As expected, multi-training and noisy outperforms the simple type of noisy training.



Results & Summary - Campus School Dataset

USING THE TIMIT TRAINED MODEL

5 Keywords - CATCH, CAT, DISH, MILK, HIDING

15 recordings all containing the 5 keywords (part of a story each ~15 seconds long) from 8 children. EDthr increased to 0.3*Len(keyword)

| **CS Dataset** | **TPR** | **FPR** |
|---|---|---|
| Wavelet MFCC(best model) | 84/530 | 446/530 |

Expected poor results because of the variations in age and accent

Other methods like Speaker adaptation and training the model on children’s speech should improve the performance of the system.

# REFERENCES

[1] Chorowski, Jan K., et al. "Attention-based models for speech recognition." Advances in neural information processing systems. 2015.

[2] https://github.com/awni/speech/tree/master/