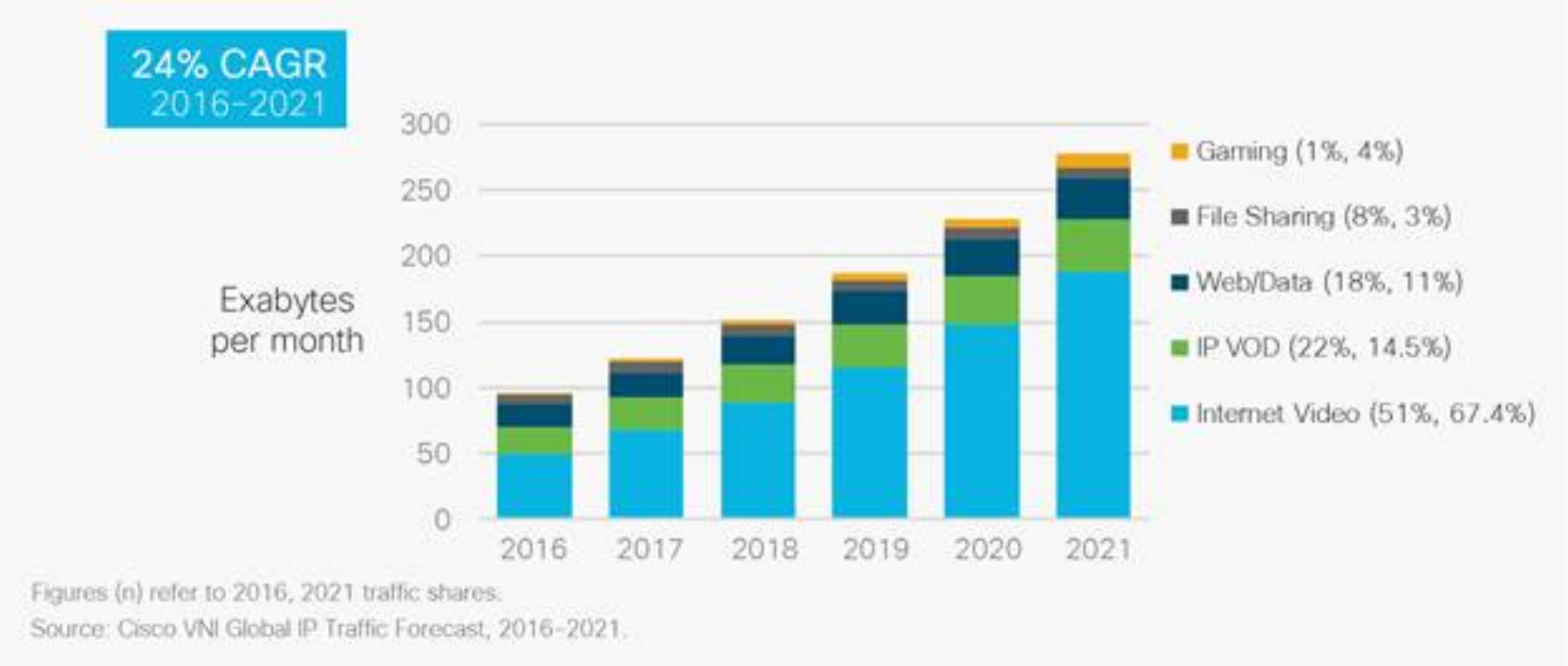
Speak search:
ASR-free search for speech corpus

Divya Pitta, Fang-Yi Chiu

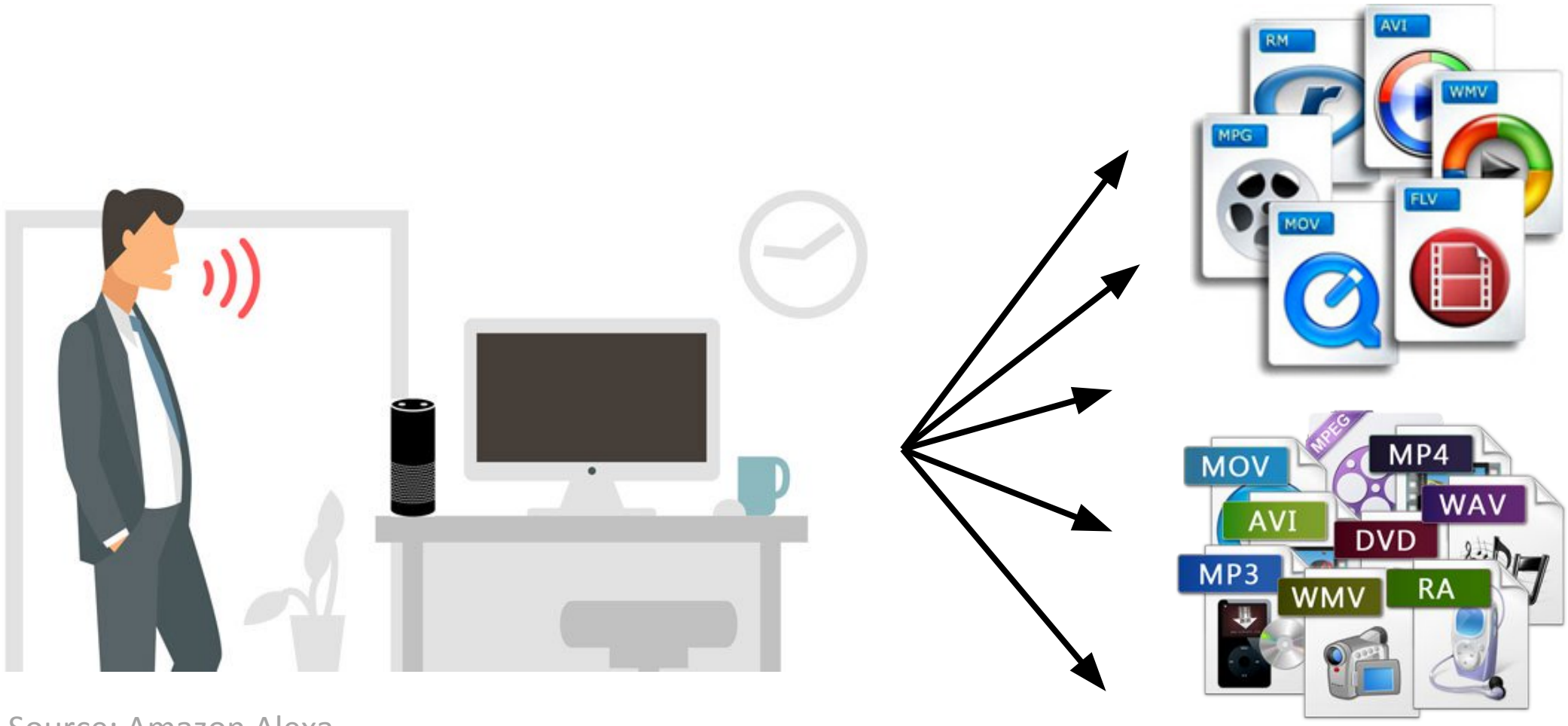
Internet Application Traffic Forecasting



Outline

- Task Scope
- Dataset Setting
- Baseline: ASR
- Speech Encoding Result
- Future Work

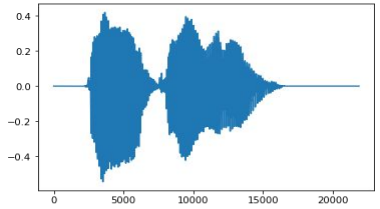
Task Scope



Source: Amazon Alexa

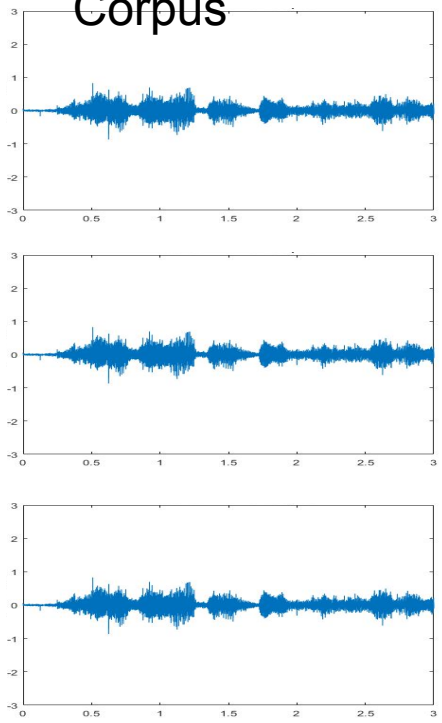
Task Scope

Query Speech

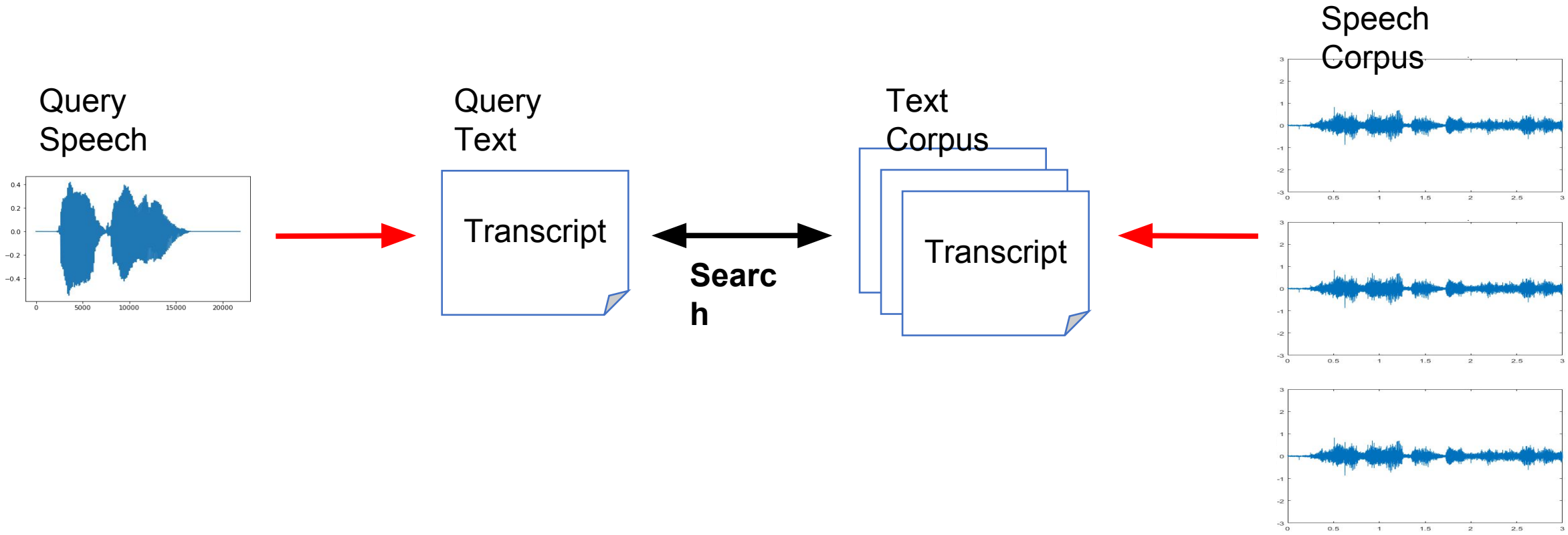


Searc h

Speech Corpus



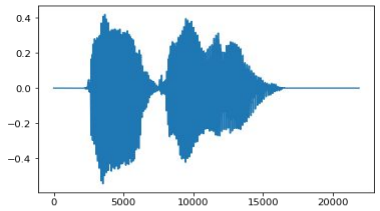
Current Solutions



Current Solutions

```
Hansel      ->      cancel          Hansel      <-      Hansel
stevie      ->      TV              stevie      <-      stevie
```

Query Speech



Query Text

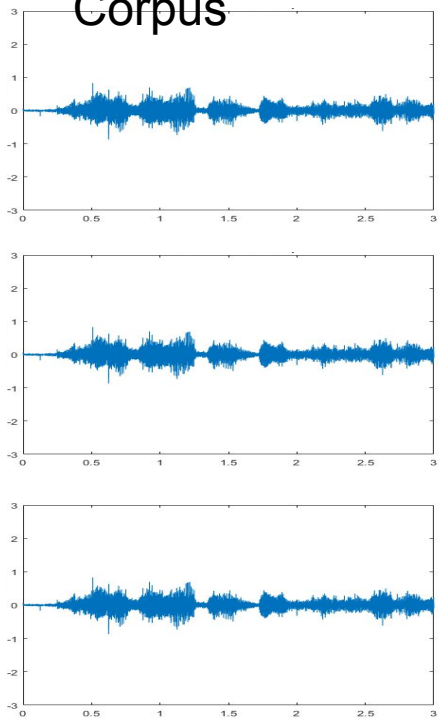
Transcript

Search

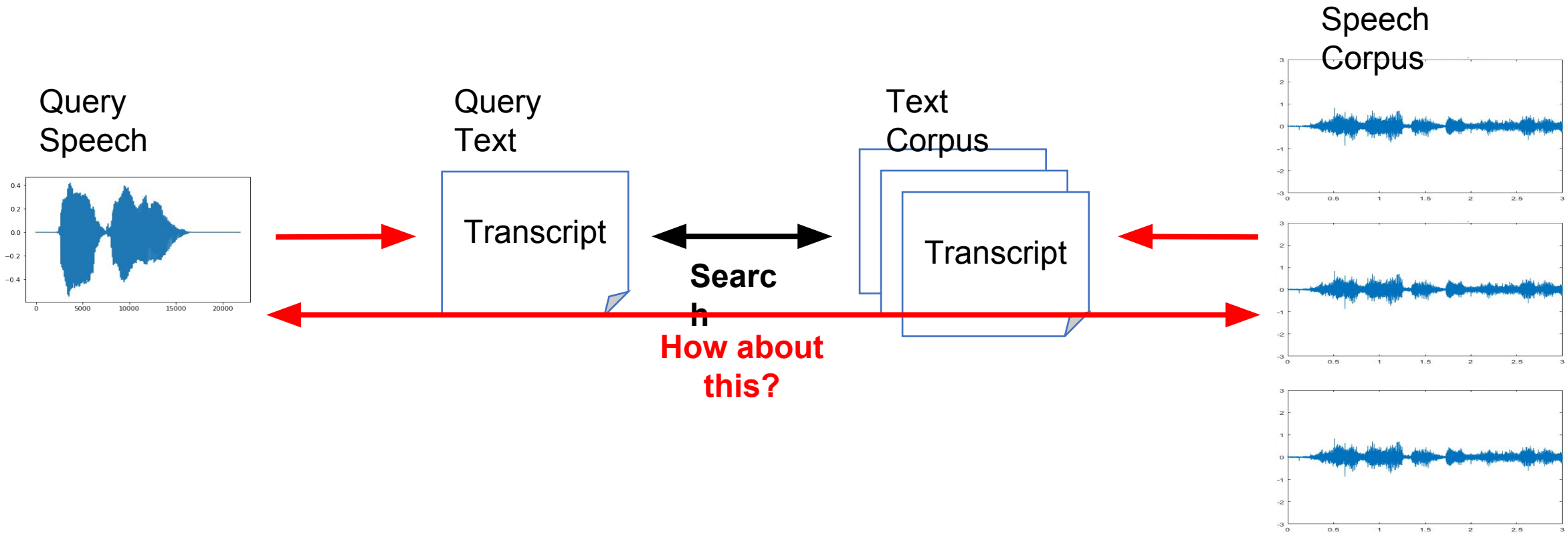
Text Corpus

Transcript

Speech Corpus



Task Scope


Query
Speech
Query
Text
Transcript
Search
How about
this?
Text
Corpus
Transcript
Speech
Corpus

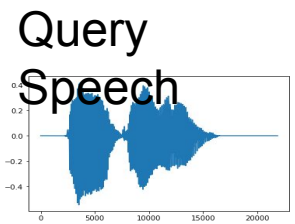
Dataset

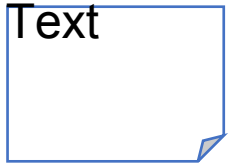
- Librispeech (dev-clean version)
  - 2700 files (337 M)
  - 2 – 33 seconds for each file
  - Transcript available for each file
  - Audio books for fiction

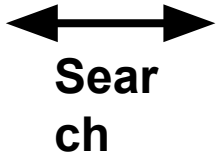
Ground Truth

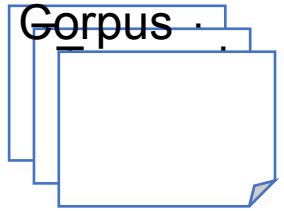
- Setting
  - Classical information retrieval problem
  - Text-to-text search
- Search engine library: Whoosh
  - Ranking by Okapi BM25 (Tf-Idf): top-k documents as ground truth
  - Default parsing: no stemming etc.
- Measurement
  - Precision
  - Recall

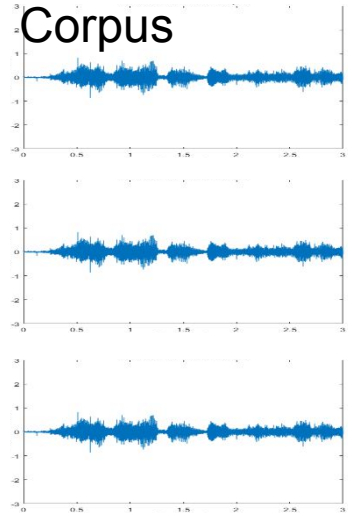
Search with ASR



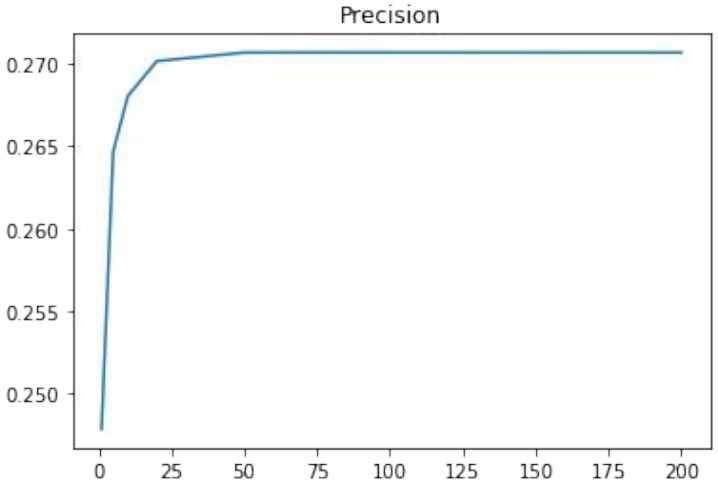


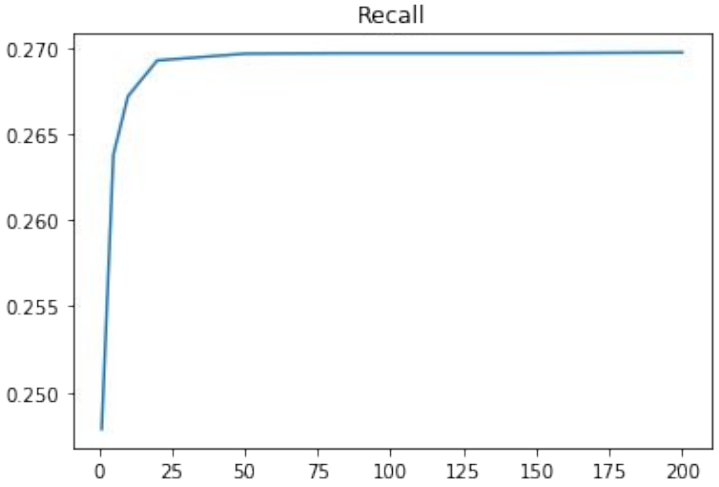


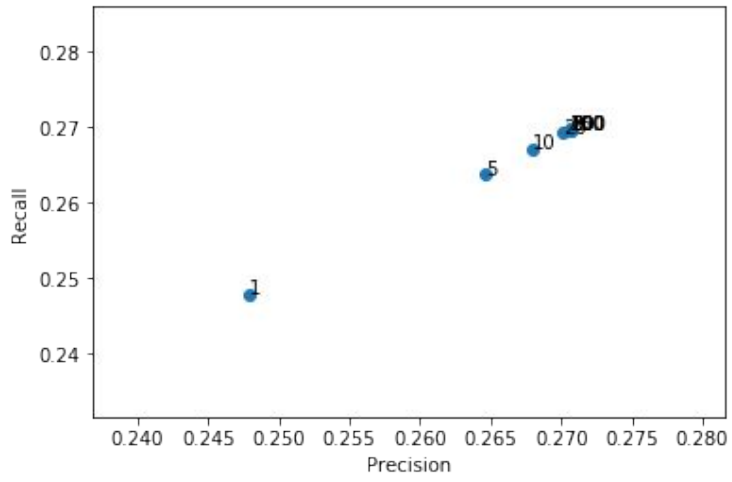




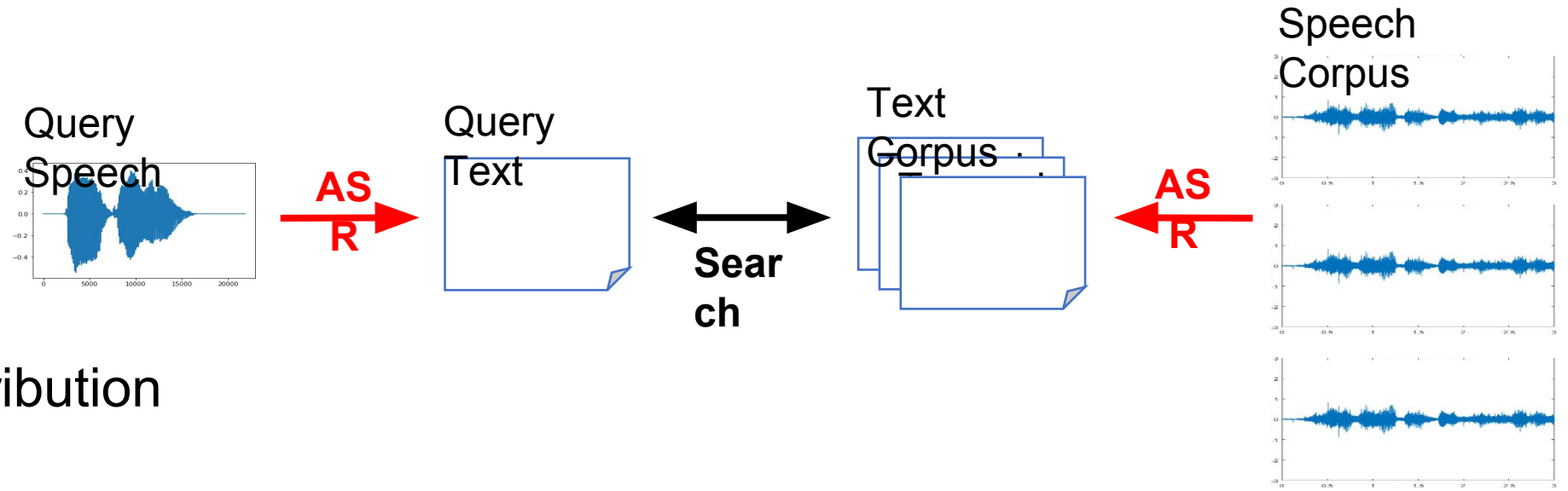
Performance: precision, recall



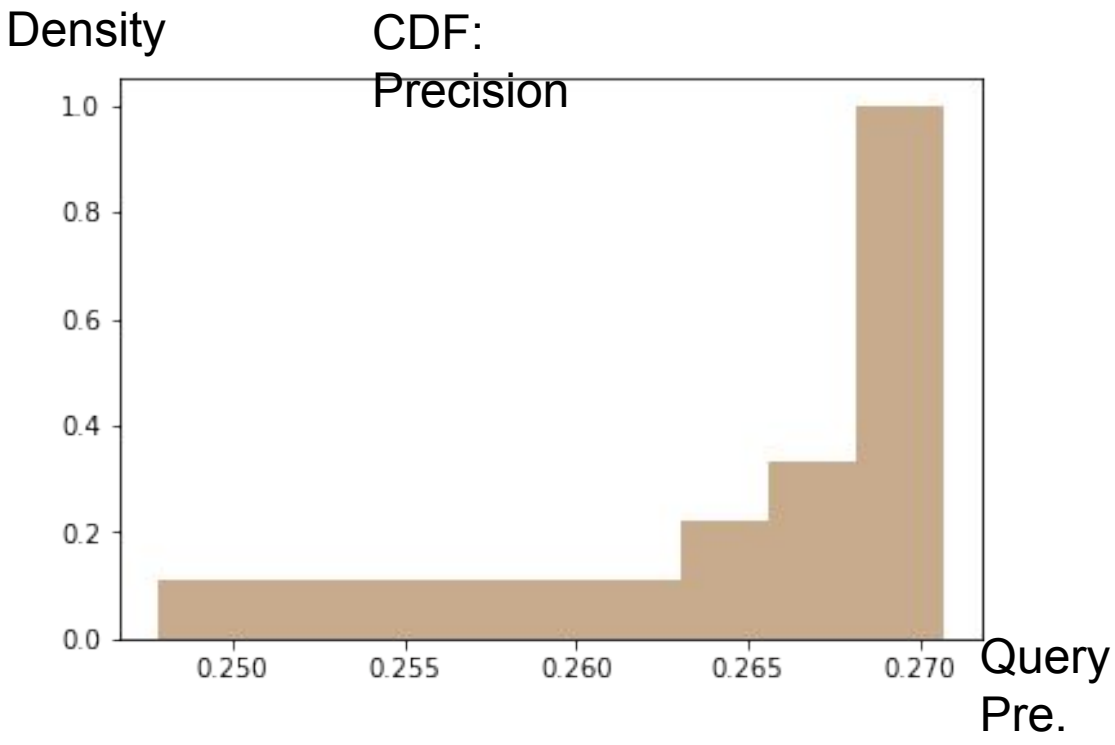


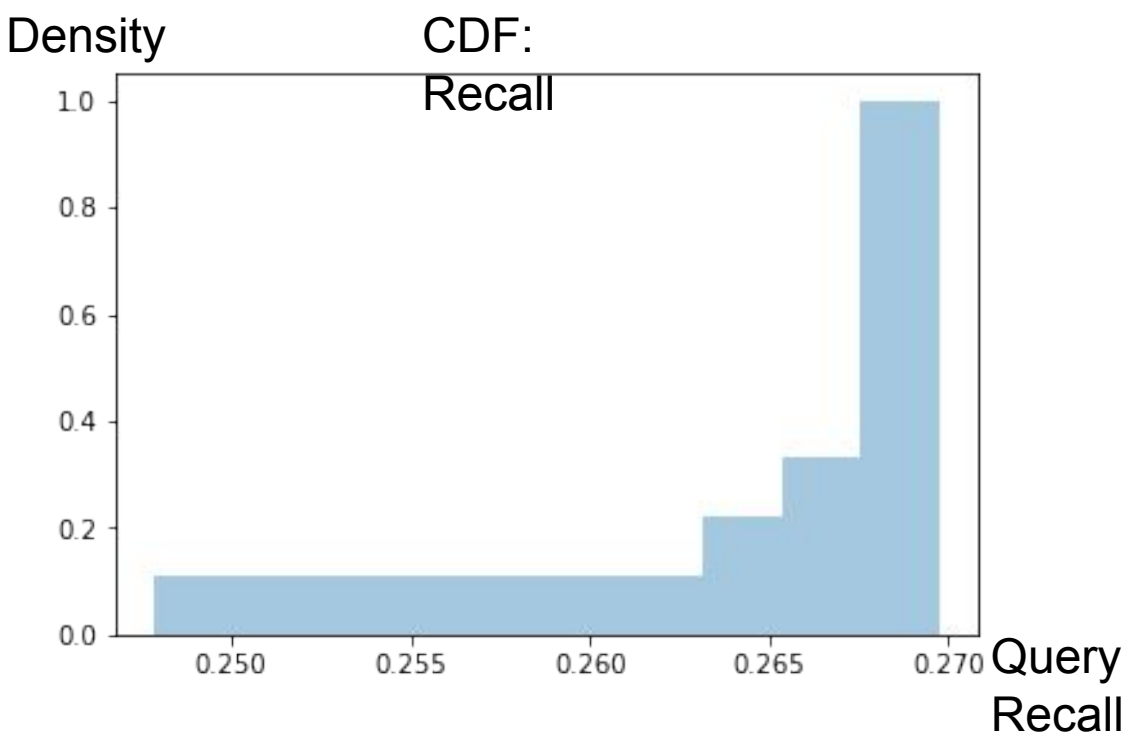


Search with ASR

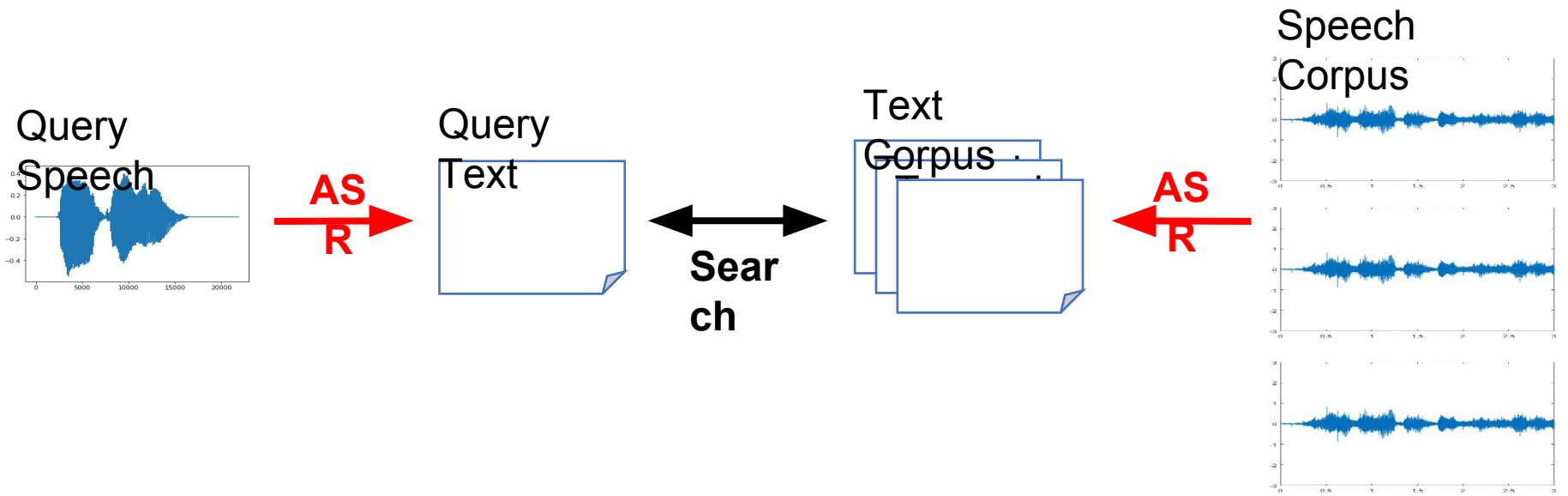


Performance distribution





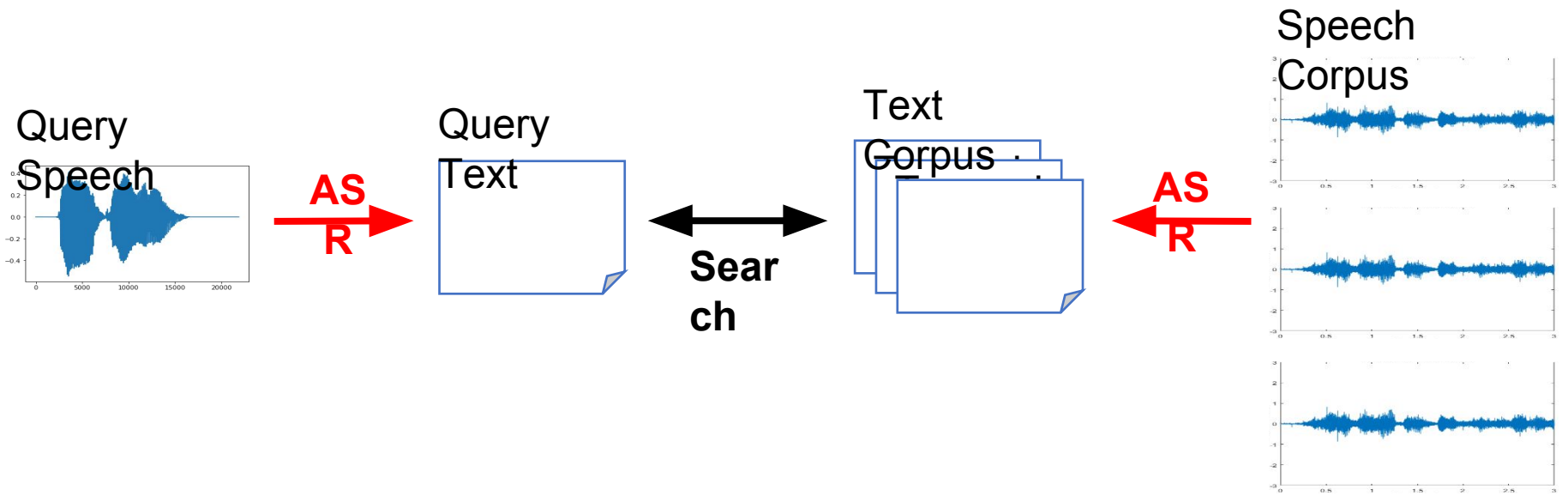
Search with ASR



- Problem
  - ASR Error: from both sided!!
  - Short token conversion
  - OOV

```
Antonia -> Anthony
ph d -> phD
rangitata (place) -> rangi
Hansel -> cancel
```

Search with ASR

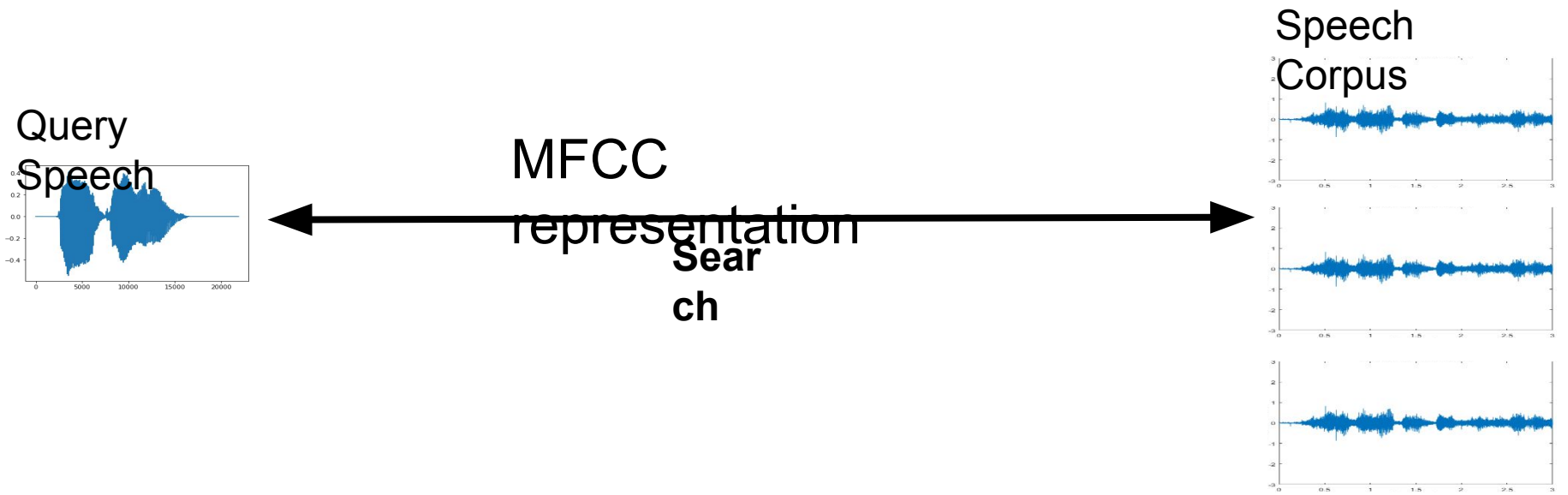


Can we reduce the errors propagated by ASR?

Task Setting

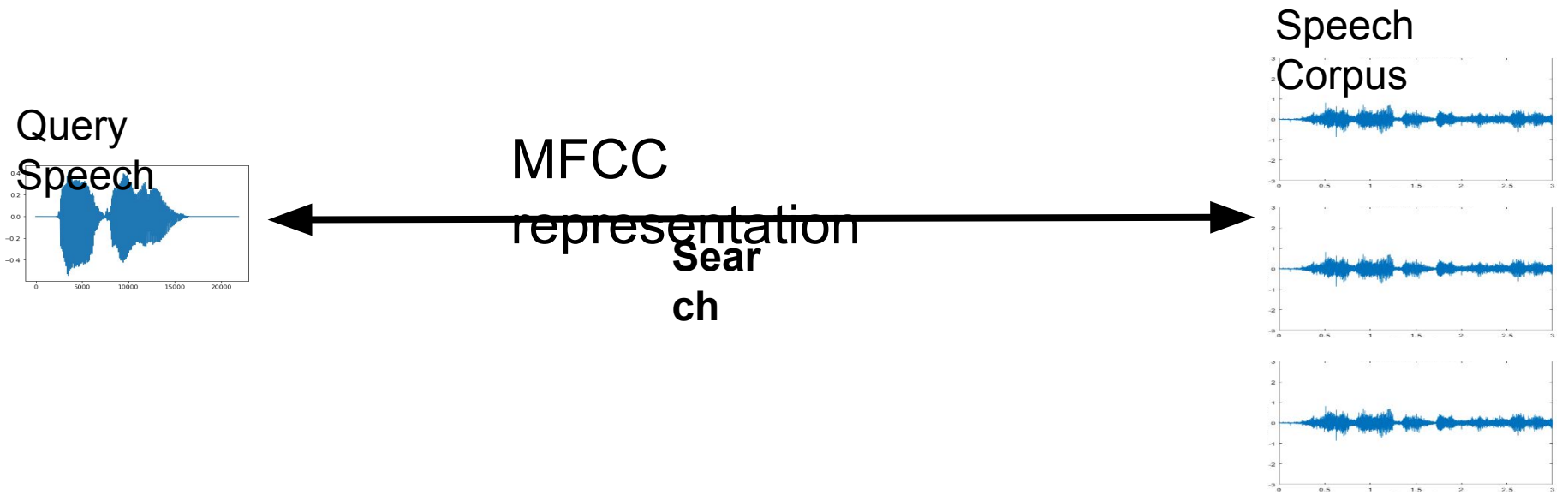
- Find embeddings of audio files
- Nearest neighbor search for each embedding

Search with MFCC



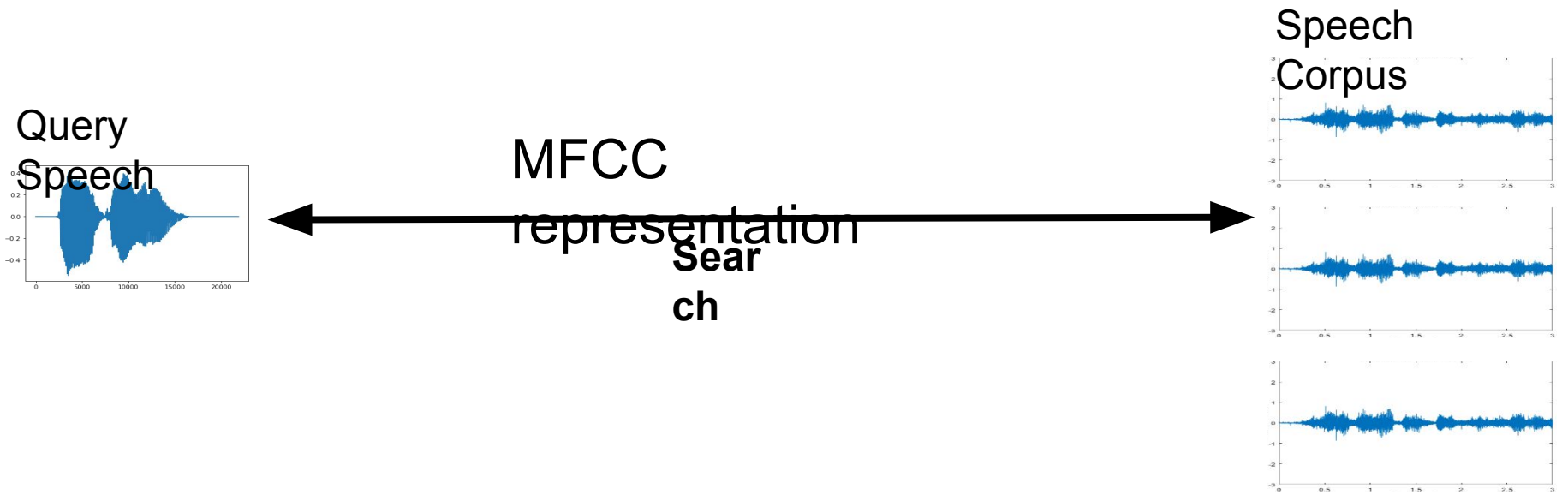
- MFCC:
  - The most common feature used in ASR system
  - Mimics people perceptron system and extract lower frequency features

Search with MFCC



- How do we search?
  - Each query has ~100 MFCC frames
  - Retrieve top-k NN for each query and then use majority vote and distance to rank

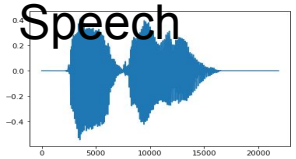
Search with MFCC



- Recall
- Precision

Search with MFCC

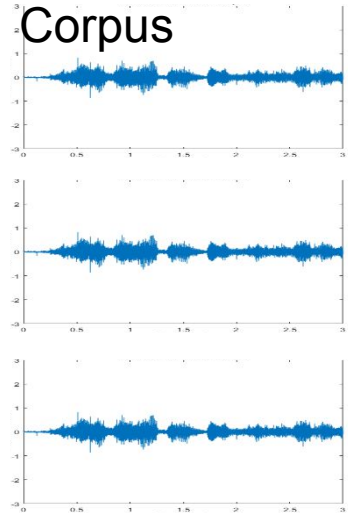
Query Speech

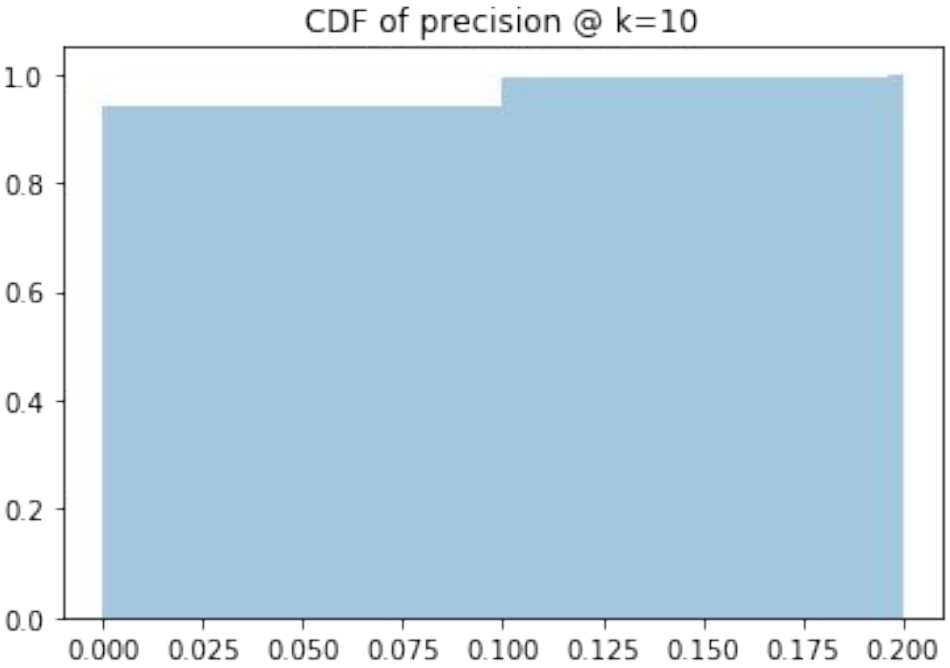


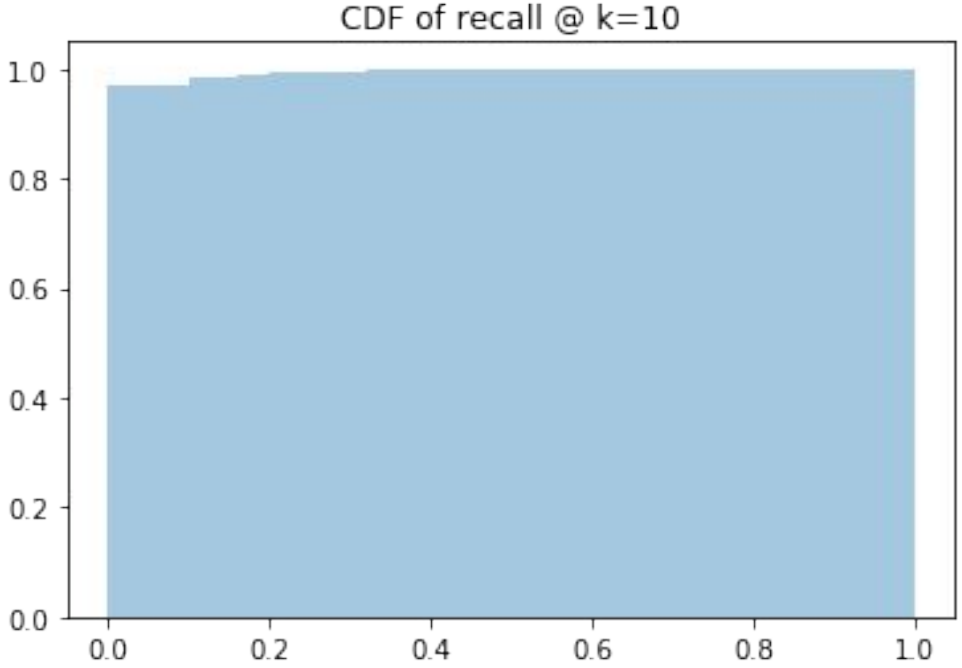
MFCC representation

Search

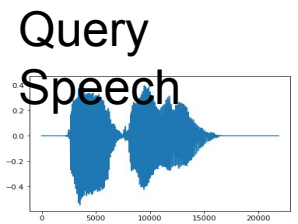
Speech Corpus





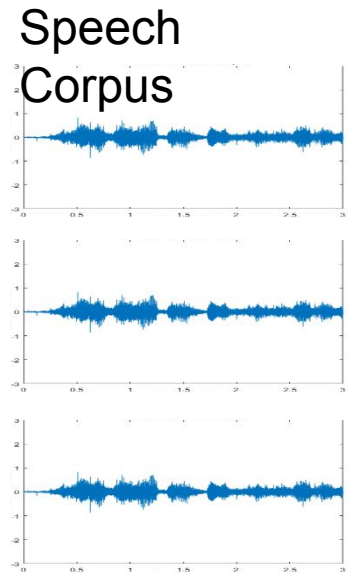


Search with MFCC

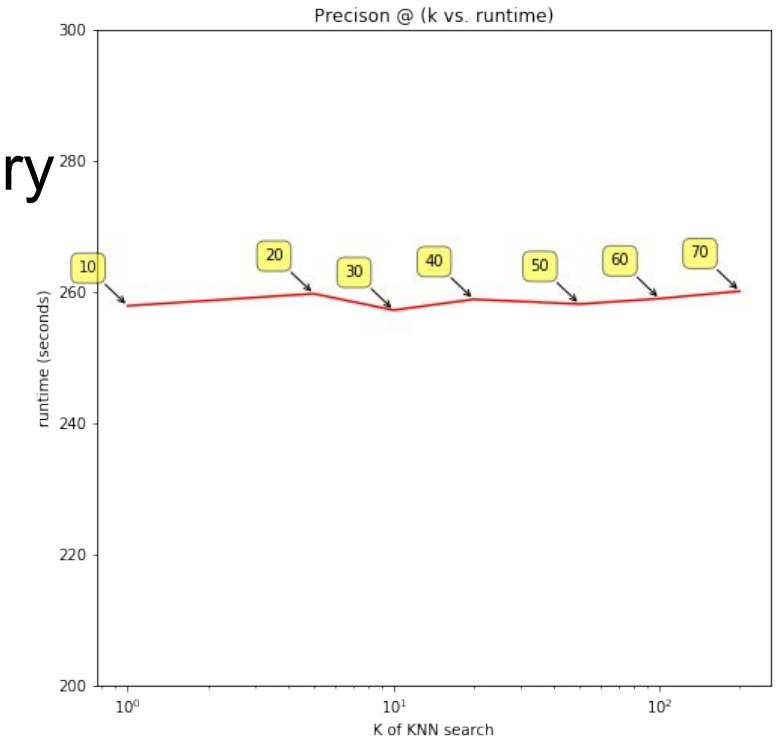


MFCC representation

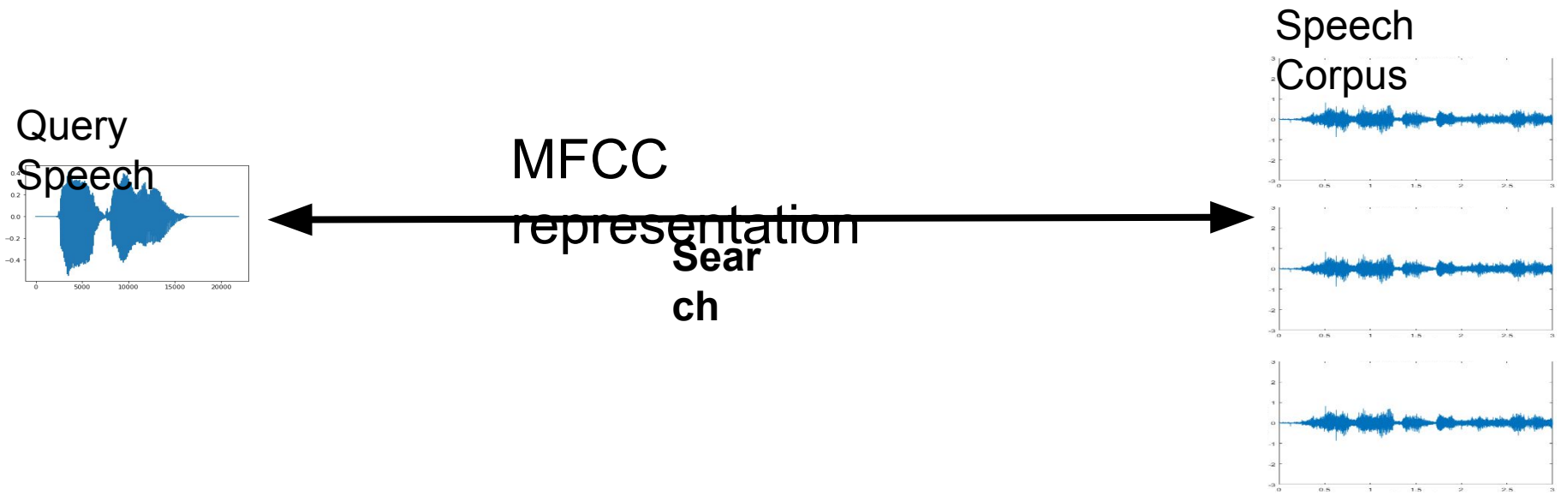
Sear ch



- Runtime analysis
  - Spark cluster with 4 workers
  - Total 32 cores and 240 G memory
  - K varies

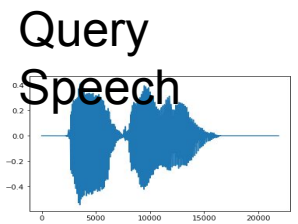


Search with MFCC



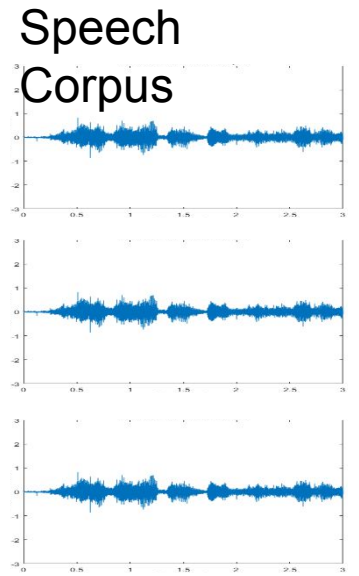
- Observation:
  - Not effective even with sliding window
  - Some queries have recall of 1.00 !!

Search with Deep Autoencoder

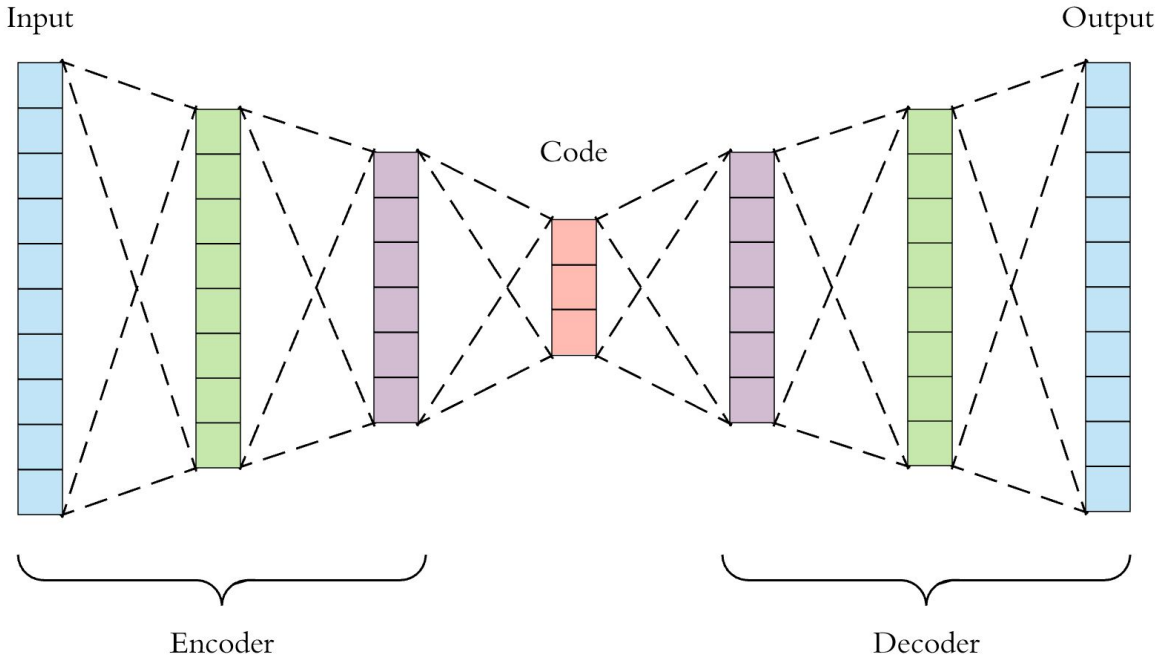


Autoencoder representation

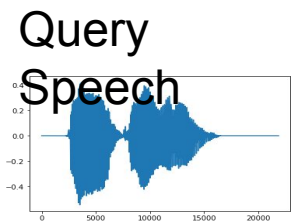
Sear ch



- Autoencoder: ANN that generates output similar to the input
  - 3 Layer: [500, 180, 120]
  - Pretrain
  - Input: 100 MFCC frames = 1200
  - Sliding windows with stride = 5

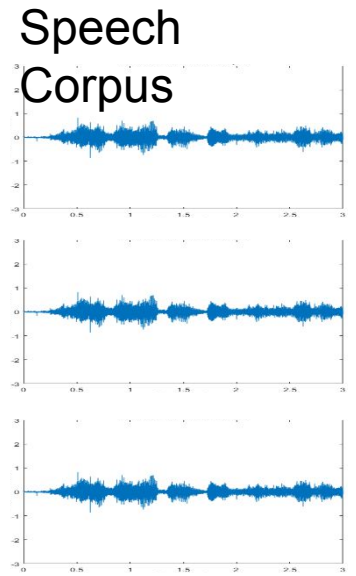


Search with Deep Autoencoder



Autoencoder representation

Sear ch



- Recall
- Precision
- Problem

Takeaway

- Signal processing is hard

Future work

- Spectrogram:
  - Time-versus-freq. features
- Last layer encoding of ASR