Speaker Change Detection

Siddharth Mittal, Neeraj Sharma, Prof. Sriram Ganapathy

Computer Science and Engineering
Indian Institute of Technology, Kanpur
Electrical Engineering
Indian Institute of Science, Bangalore

July 7, 2017

Speaker Change Detection

Given a wave file, the aim is to find out instances of speaker change.

Motivation

- Diarization: It is the task of determining who spoke where (and what). Speaker change detection is preliminary processing step for diarization.
- Also, essential in applications like conference and meeting audio data indexing.

Literature Review

- Distance metrics based classification: Using a pair of sliding windows and computing the distance between their contents.
- Build speaker models: Identify each speaker accurately then instances of speaker ID change imply a change in speaker.

Issues with current methods

- Since we have no prior knowledge of speakers, there is no data to obtain an accurate speaker model a priori.
- For the system to be real time, we can't use data hungry clustering methods like GMM.

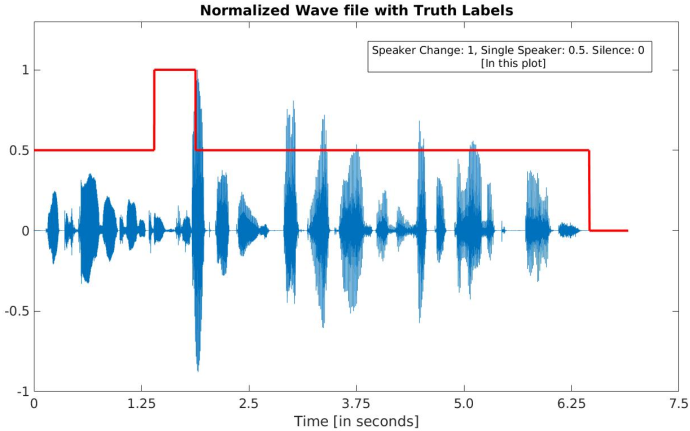
Dataset

- We used the TIMIT dataset, which consists of data from 630 speakers.
- From each speaker, we have a set of 10 sentences, and their corresponding phonetic transcriptions.
- Out of 10 sentences, 2 sentences are common to every speaker, 3 are unique to the speaker and other 5 are spoken by 7 speakers each.

Data Preparation

- We create three disjoint list of speakers, corresponding to training, validation and test sets. For each speaker we only use the diverse (spoken by individual) and compact (spoken by 7 speakers).
- First, from the phonetic transcriptions we find the speech regions of two random wave files.
- Since the wave files have some silence at the start and end of file, we concatenate the two wave files in a way ensuring there is no silence between the two speakers.

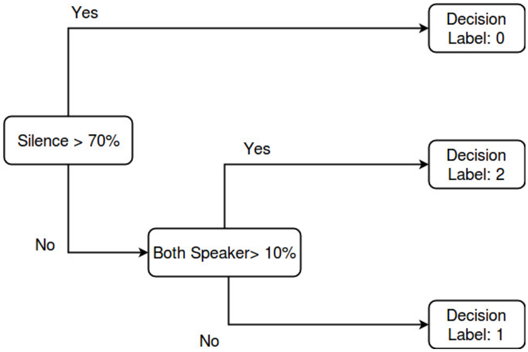
Wave file with truth labels



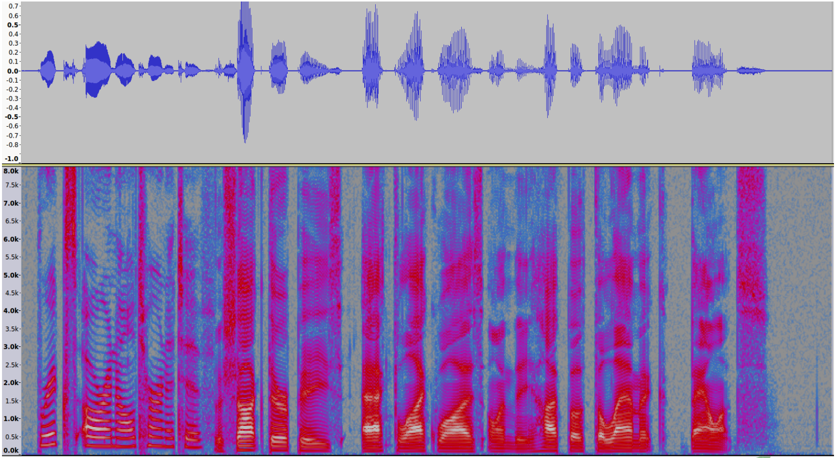
Labelling strategy

- The main issue in speaker change problem is how to obtain the ground truth.
- Since we always consider a non-zero length window to compute any feature, marking the change as instantaneous won't help.
- To take decision whether a frame is speaker change or not, we consider windows of 200 msec, 400 msec and 600 msec.

Labelling Strategy



Spectrogram and Wave File



Features and Approaches Tried

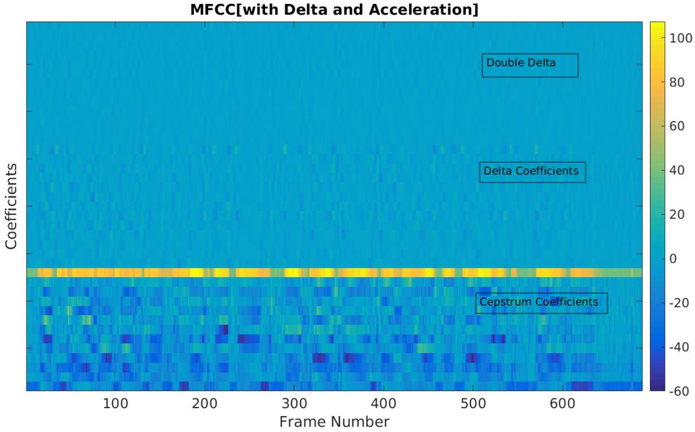
- The idea was to capture the subtle changes in the features, that occur when the speaker change occurs.
- Choosing the right window for decision is critical for detecting the change.

MFCC

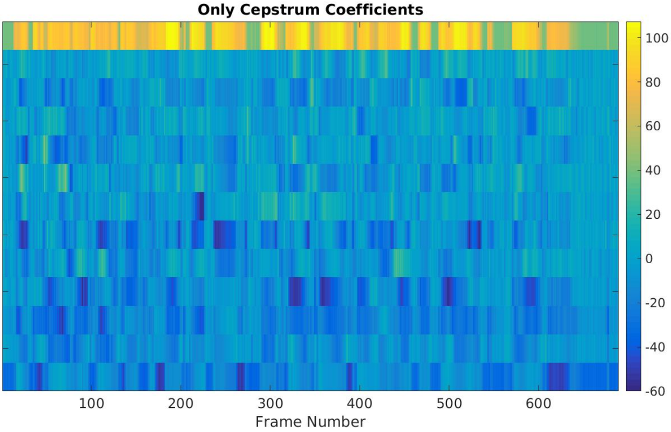
Steps to obtain MFCC:

- Frame the signal into short frames.
- For each frame calculate the periodogram estimate of the power spectrum.
- Apply the mel filterbank to the power spectra, sum the energy in each filter.
- Take the logarithm of all filterbank energies.
- Take the DCT of the log filterbank energies.
- Keep DCT coefficients 1-13, discard the rest.

MFCC-Delta-Acceleration

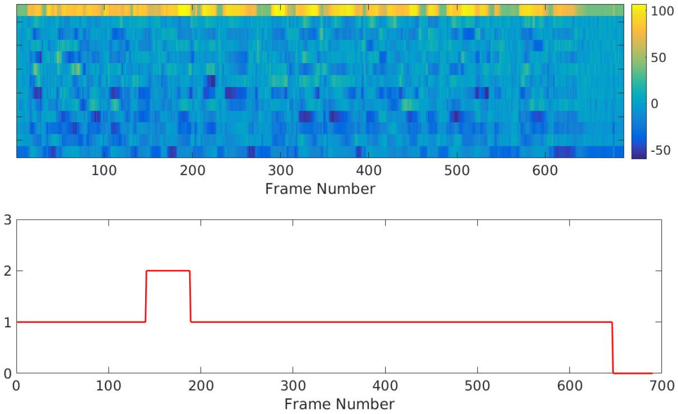


Only MFCC

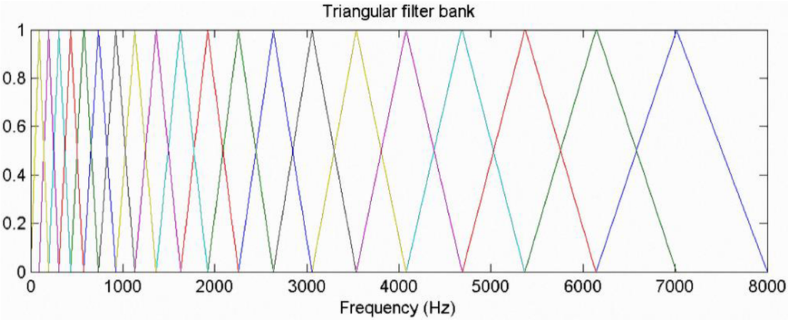


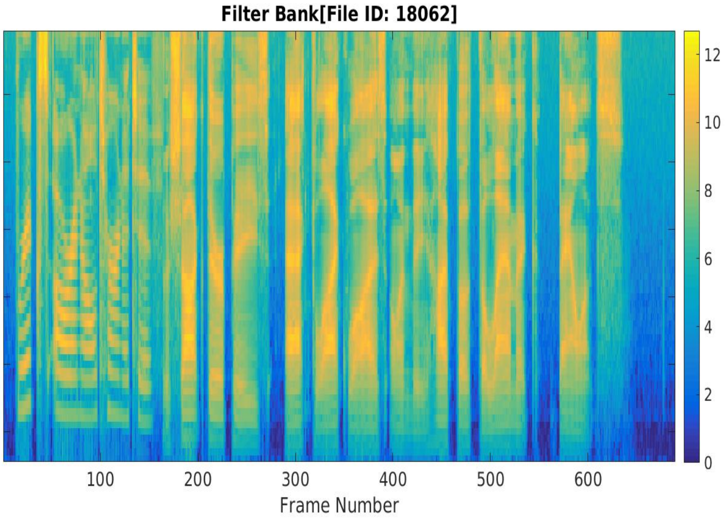
Only Cepstrum Coefficients
100
80
60
40
20
0
-20
-40
-60
100
200
300
400
500
600
Frame Number

MFCC with truth label



Mel Filter Bank

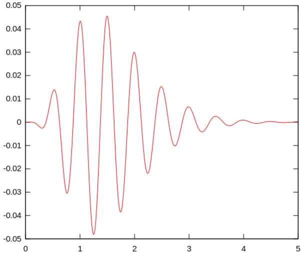



Filter Bank[File ID: 18062]
100
200
300
400
500
600
Frame Number
0
2
4
6
8
10
12

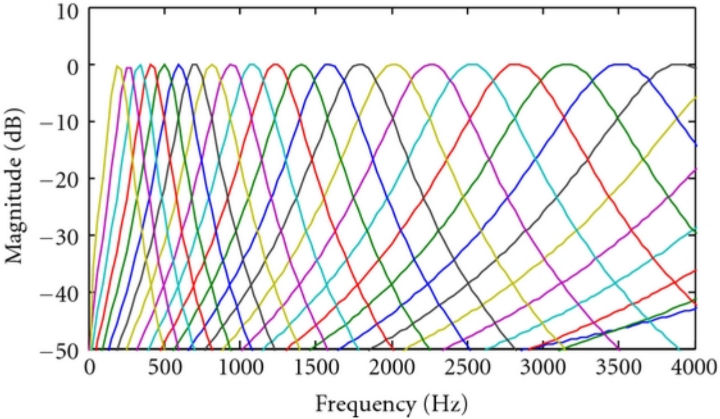
Gammatonegram

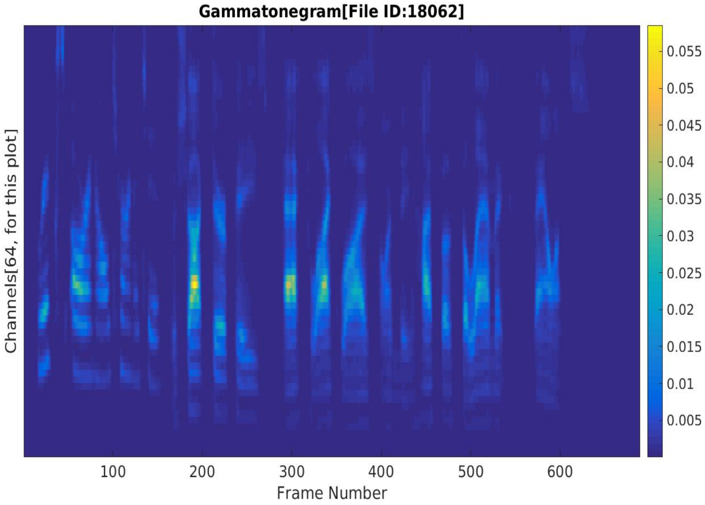
The gammatone filter models:

$$g(t) = at^{n-1} \exp^{-2\pi bt} \cos(2\pi ft + \phi)$$

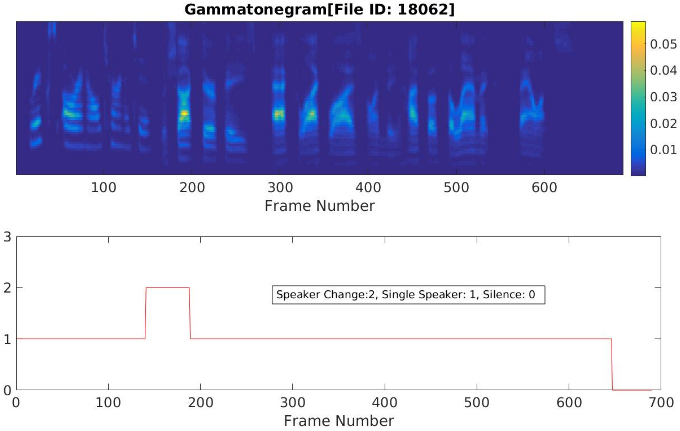


Gammatone Filterbank frequency response

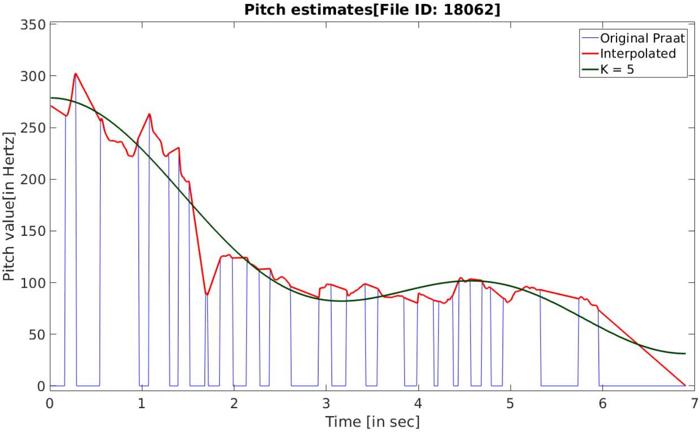



Gammatonegram[File ID:18062]
Channels[64, for this plot]
Frame Number
100
200
300
400
500
600
0.055
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005

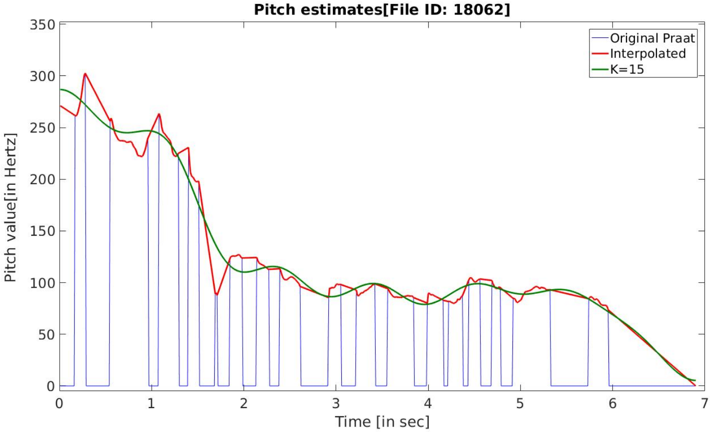
Gamma features with truth labels



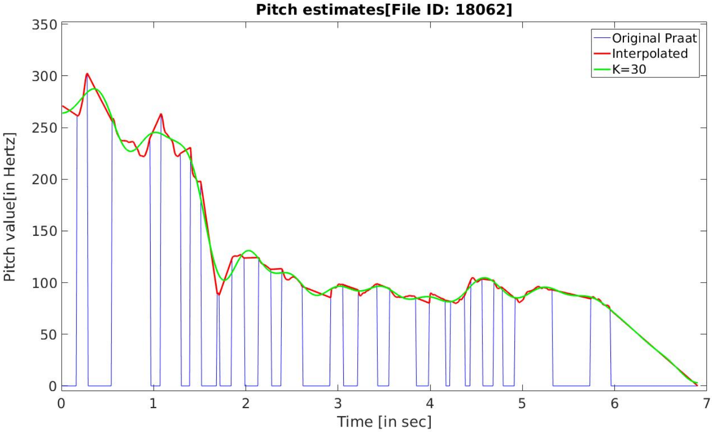
Pitch estimates (K=5)



Pitch estimates (K=15)



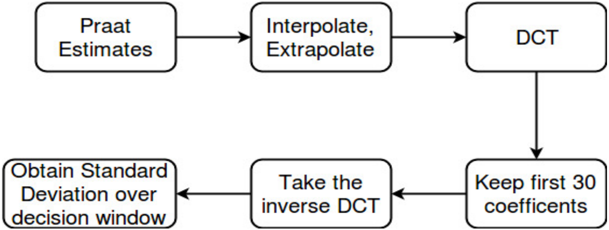
Pitch estimates (K=30)



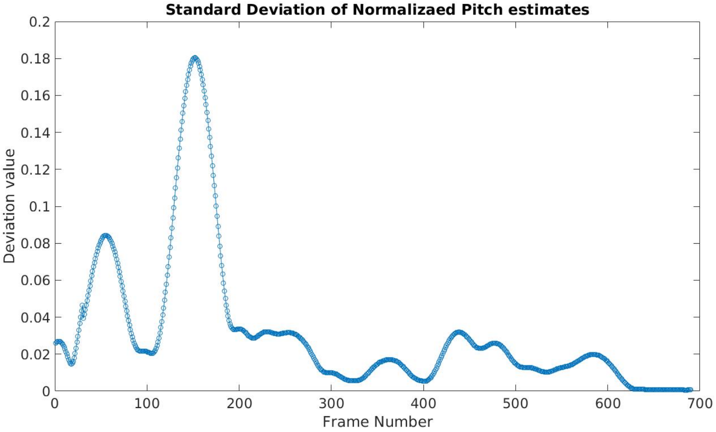
Steps used to obtain Pitch estimates

- First, using Praat pitch estimates are obtained for the given file, for a window of 25ms with a hop of 10ms
- For the regions where Praat can't estimate the pitch value, it gives "-undefined–" as its output. We change this undefined to zero.
- Next, we use linear interpolation and extrapolation on non-zero points, to obtain a continuous pitch estimate for the whole signal
- We take the DCT of the estimate, and zero out coefficients greater than 30, and take i-DCT to get smoothened pitch estimates
- Next, we take the standard deviation of the above obtained pitch estimates, over a window of 610 ms, with a hop of 10 ms.

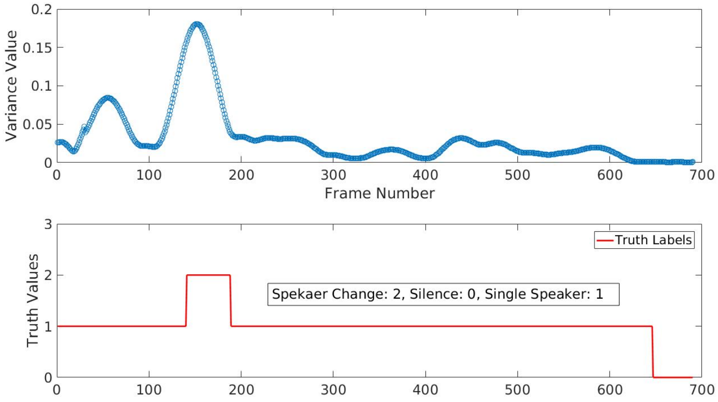
Steps Involved

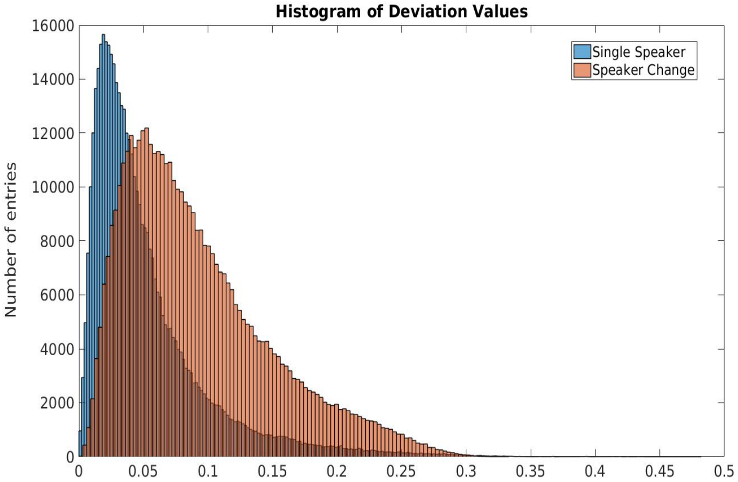


Deviation of Pitch estimates



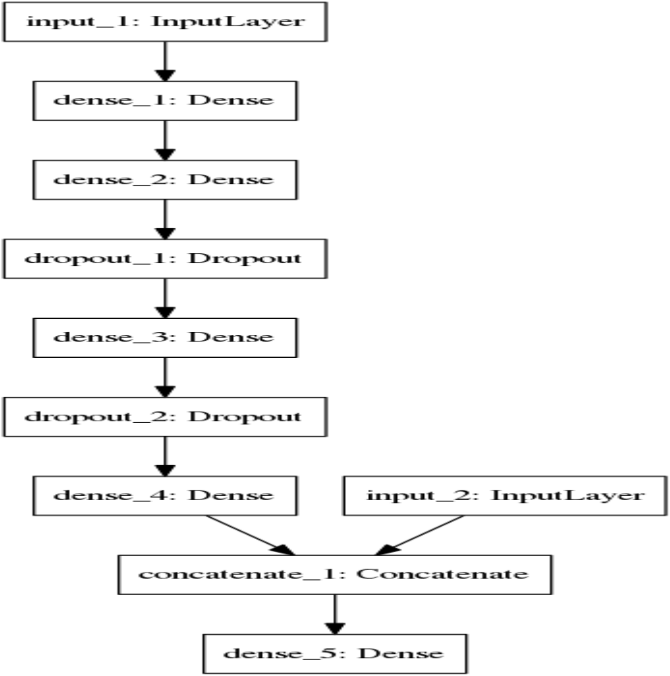
Deviation with ground truth labels




Histogram of Deviation Values
Single Speaker
Speaker Change
Number of entries
0
2000
4000
6000
8000
10000
12000
14000
16000
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5

Results[with DNN Classifier]

Context	Features	1 Speaker[Val Acc.]	Sp. Change[Val Acc.]	F-F	M-M	F-M	%SC[Train data]
200 msec	MFCC Filter Bank Gammatone	Features not good					
400 msec	MFCC Filter Bank Gammatone	Features not good					
600 msec	MFCC Filter Bank	Features not good					
	Gammatone	85%	82%	83%	76%	84%	48%
	Gammatone-Pitch	73%	91.3%	88.7%	89.29%	93.38%	54%
800 msec	Gammatone	89%	70%	71%	58%	73%	63.5%
	Gammatone-Pitch	72.68%	88.84%	88.2%	84.79%	91.56%	55.33%


input_1: InputLayer
dense_1: Dense
dense_2: Dense
dropout_1: Dropout
dense_3: Dense
dropout_2: Dropout
dense_4: Dense
input_2: InputLayer
concatenate_1: Concatenate
dense_5: Dense

Some key findings

- As expected, female-male speaker changes have accuracy's higher than female-female and male-male.
- On including pitch estimates, the accuracy of Male-Male speaker change detection increased.
- The models which use pitch estimates[deviations] generalize better than those models which don't include pitch estimates.

Future work

- Assigning a confidence measure to frames, so as to account for number of samples from speaker 1 and for speaker 2, for the given decision window.
- Using Siamese network to compare between different features

Thank You.

The scripts used, and other codes used can be found at:
`https://github.com/smittal6/leap-scd`