Talgreining á íslensku

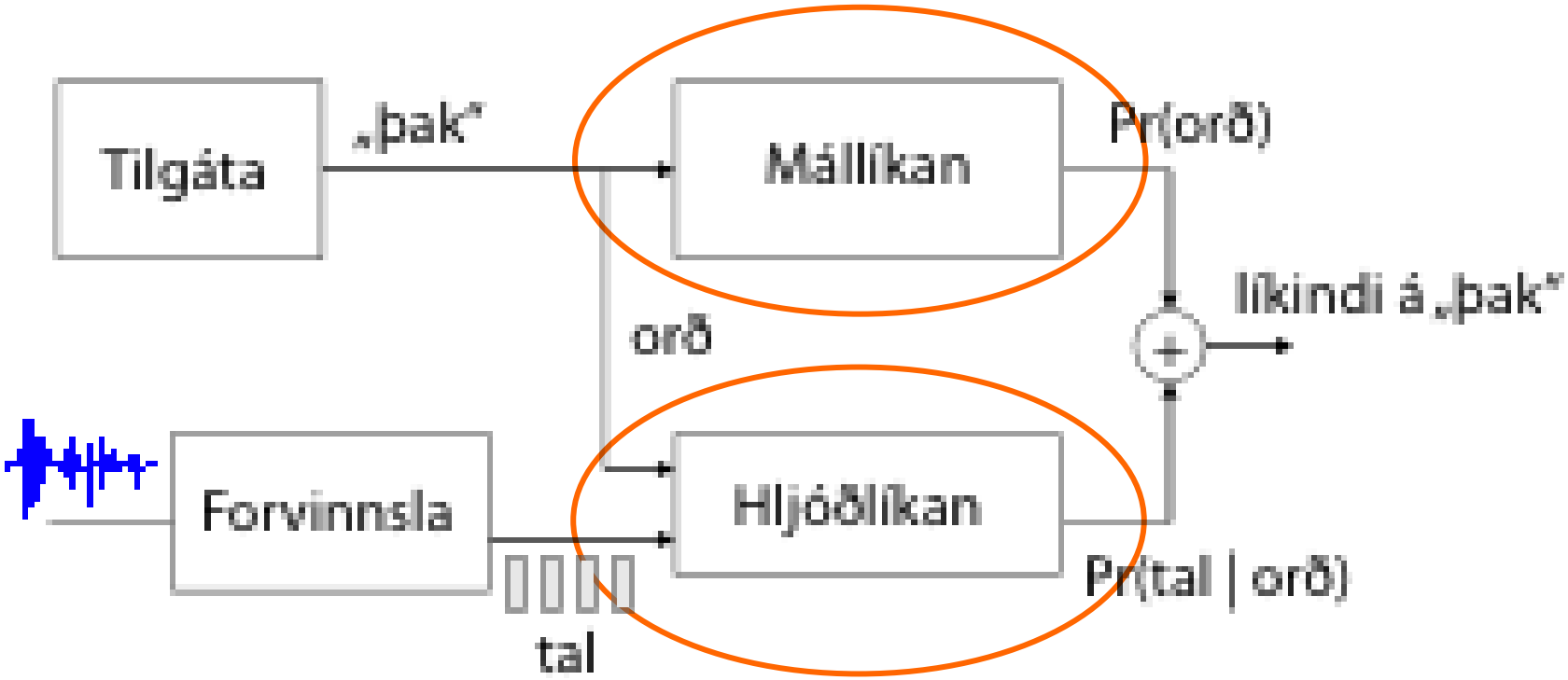
Er samstarf við Google nóg
eða eigum við að ráða yfir þessari tækni sjálf?

JÓN GUÐNASON
HÁSKÓLINN Í REYKJAVÍK

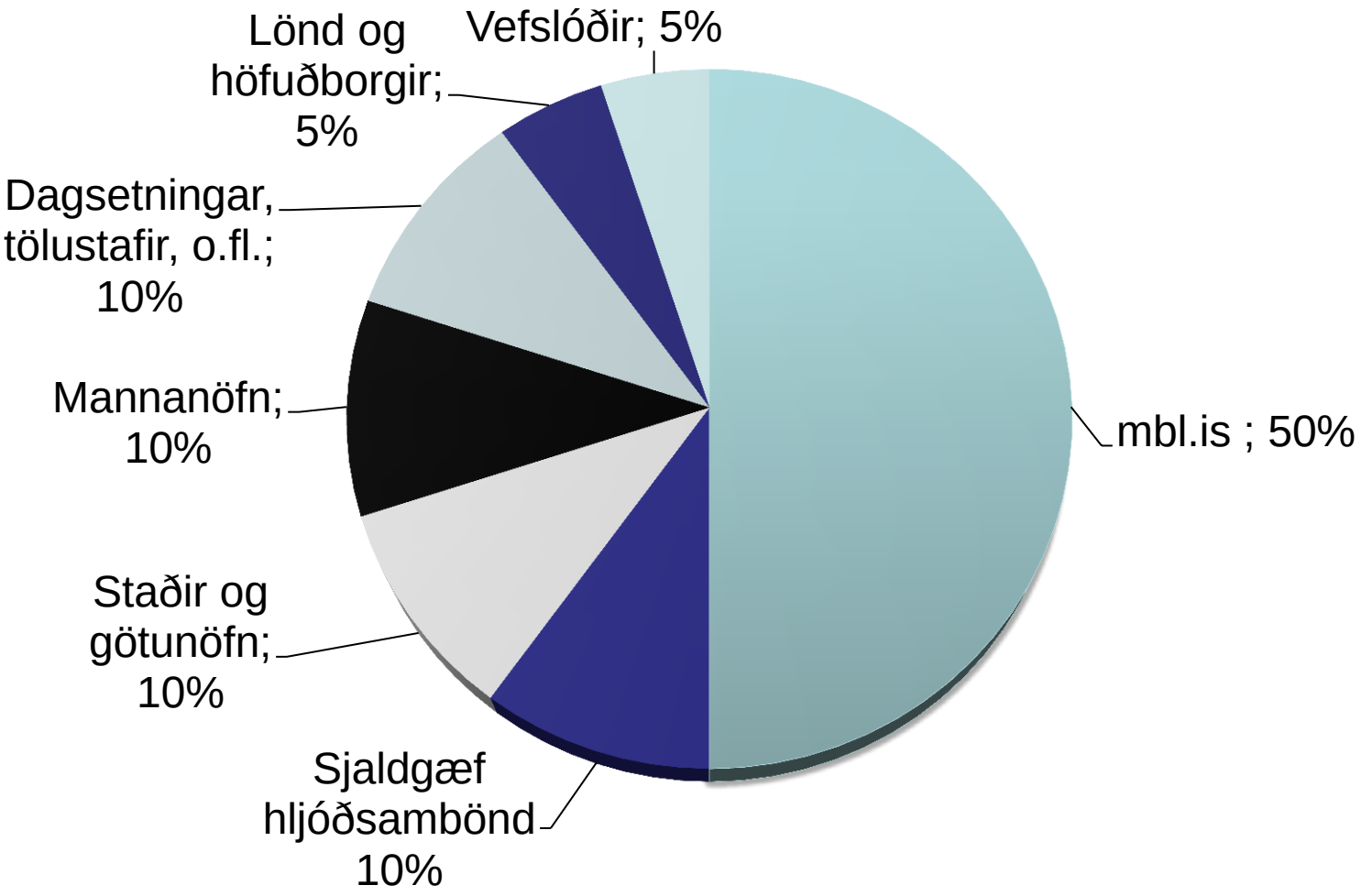
Yfirlit

- Gagnasöfnun og Google verkefni
- Talgreining
  - Mállíkan
  - Hljóðlíkan
  - Merkjavinnsla
- Hvað þarf til þess að þróa talgreini fyrir íslensku
- Tækni sem býður upp á möguleika fyrir íslenskt samfélag og atvinnulíf

Talgreining og gagnaöflun



Söfnun gagna í samstarfi við Google



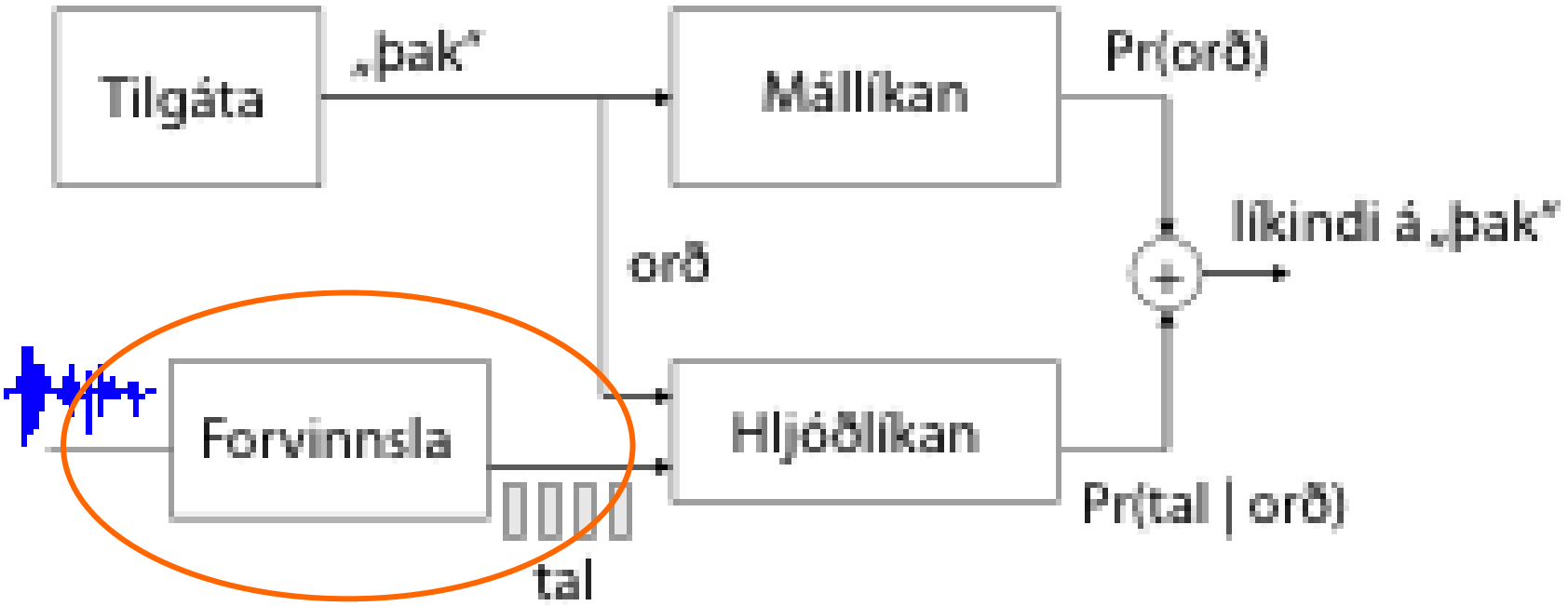
Söfnun gagna í samstarfi við Google

- 10 Android G1 símar notaðir við upptökur
- Hver notandi las 200-500 setningar
- Setningalistinn geymdur á Google netþjóni
- Við hvern upplestur er hljóðskrá send á neþjón og geymd með texta

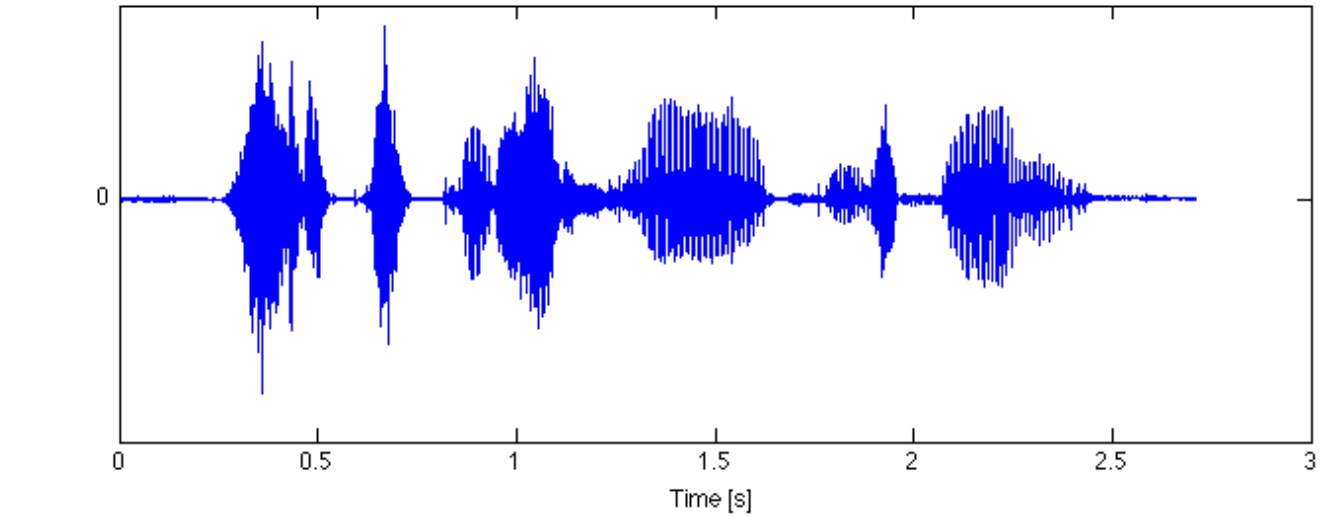
Samsetning gagnasafnsins

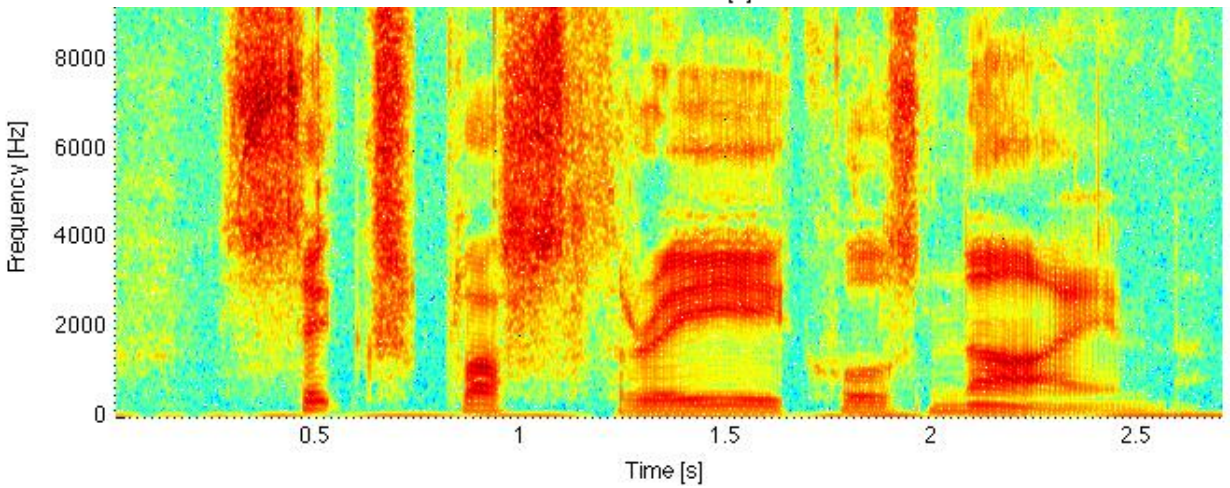
	Karlar	**Konur**	**Heild**
Þátttakendur	303 (53,8%)	260 (46,2%)	**563**
Setningar	63.215 (51,3%)	60.012 (48,7%)	**123.227**
Setn./Þáttt.	208,6	230,8	**218,9**

Talgreining og gagnaöflun

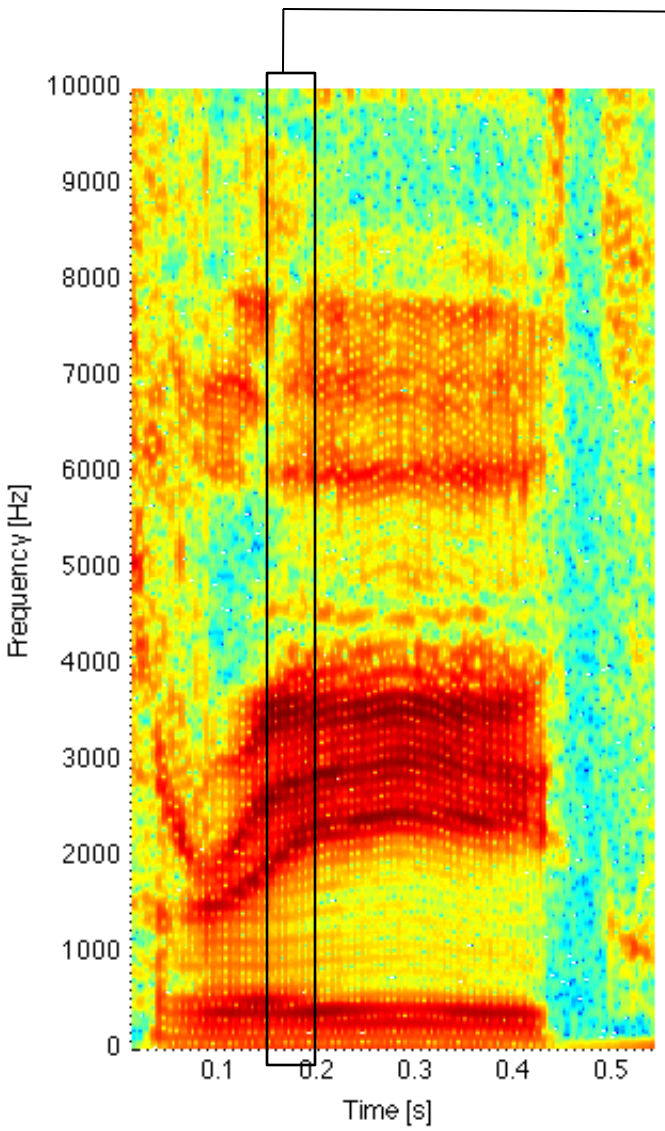


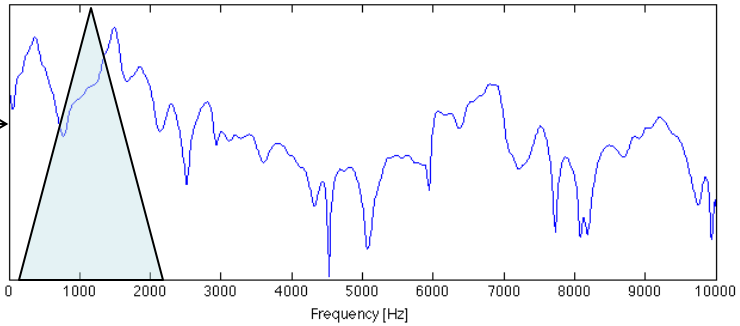
Tal sem hljóðmerki

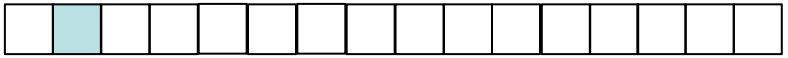




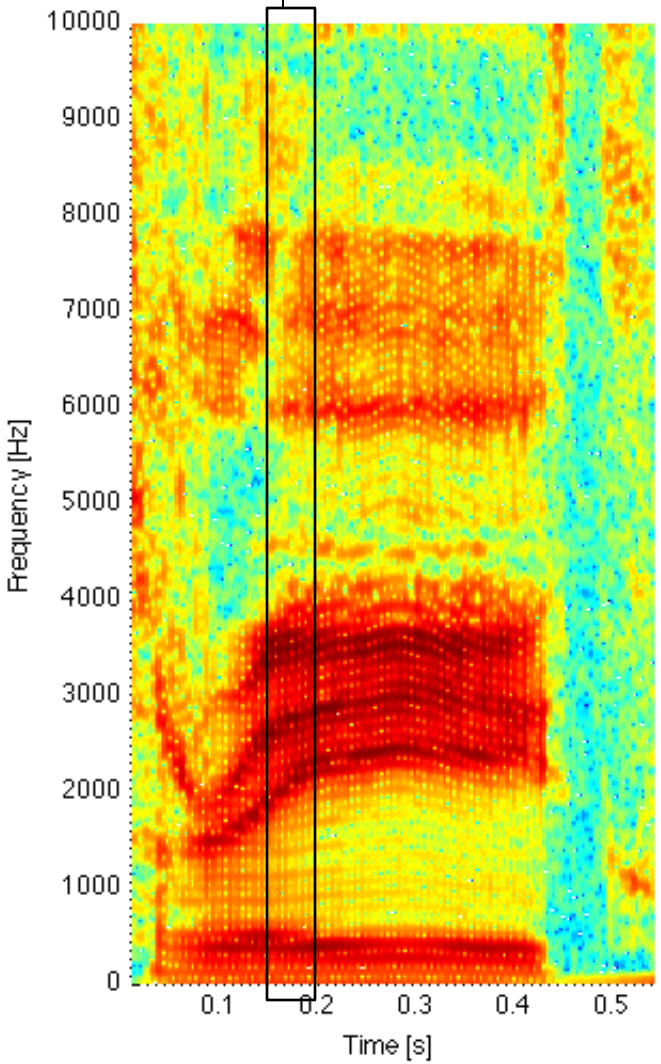
Forvinnsla

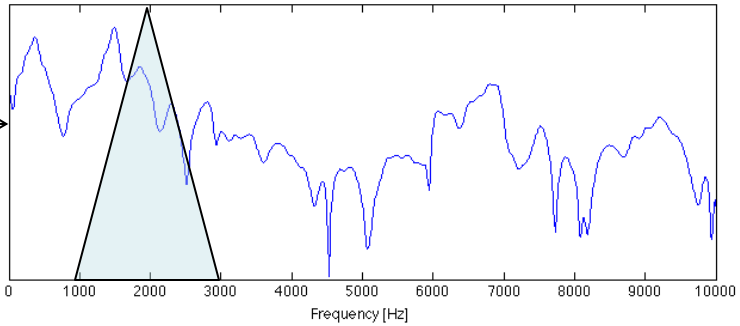

10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency [Hz]
0.1
0.2
0.3
0.4
0.5
Time [s]

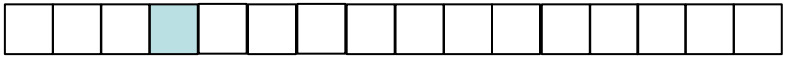

0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Frequency [Hz]



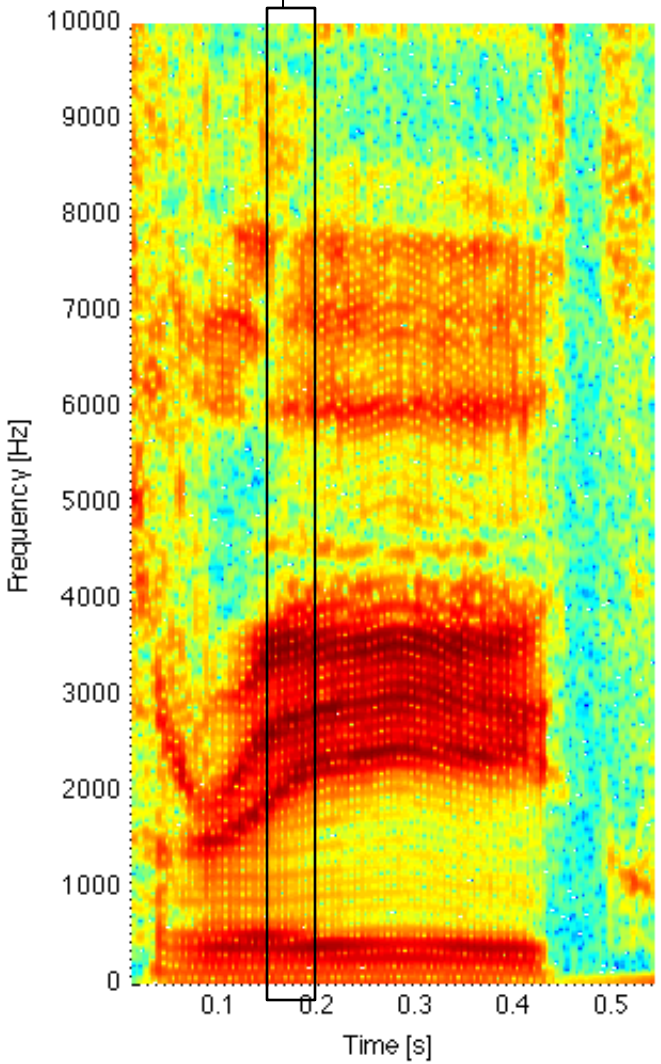
Forvinnsla

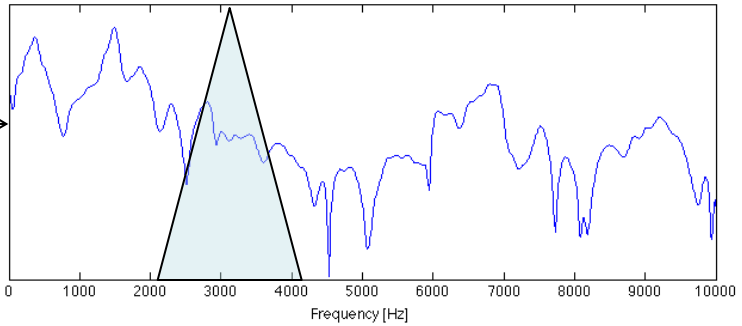

10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency [Hz]
0.1
0.2
0.3
0.4
0.5
Time [s]

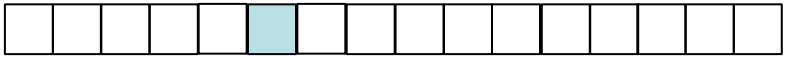

0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Frequency [Hz]



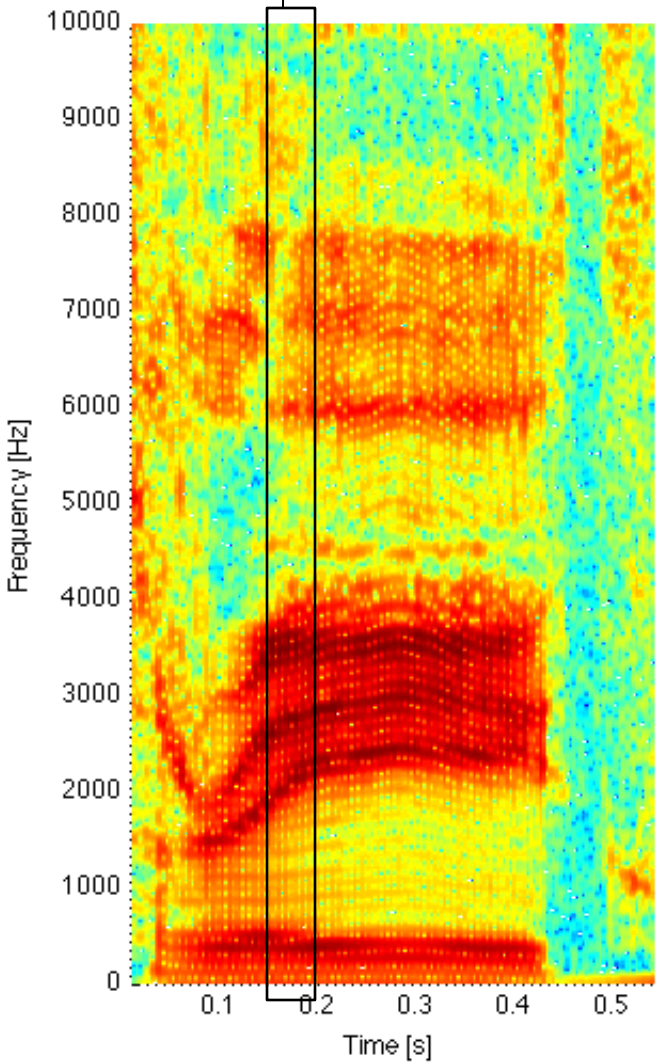
Forvinnsla

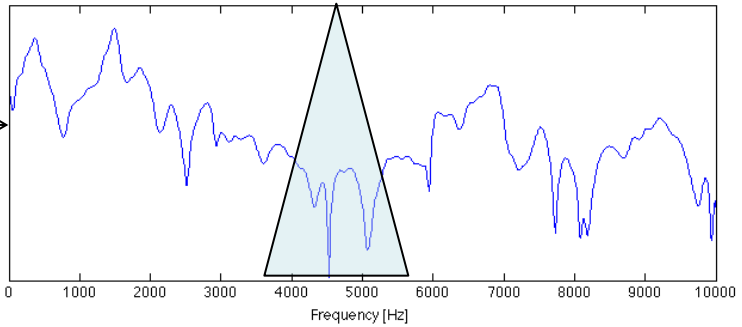

10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency [Hz]
0.1
0.2
0.3
0.4
0.5
Time [s]


0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Frequency [Hz]

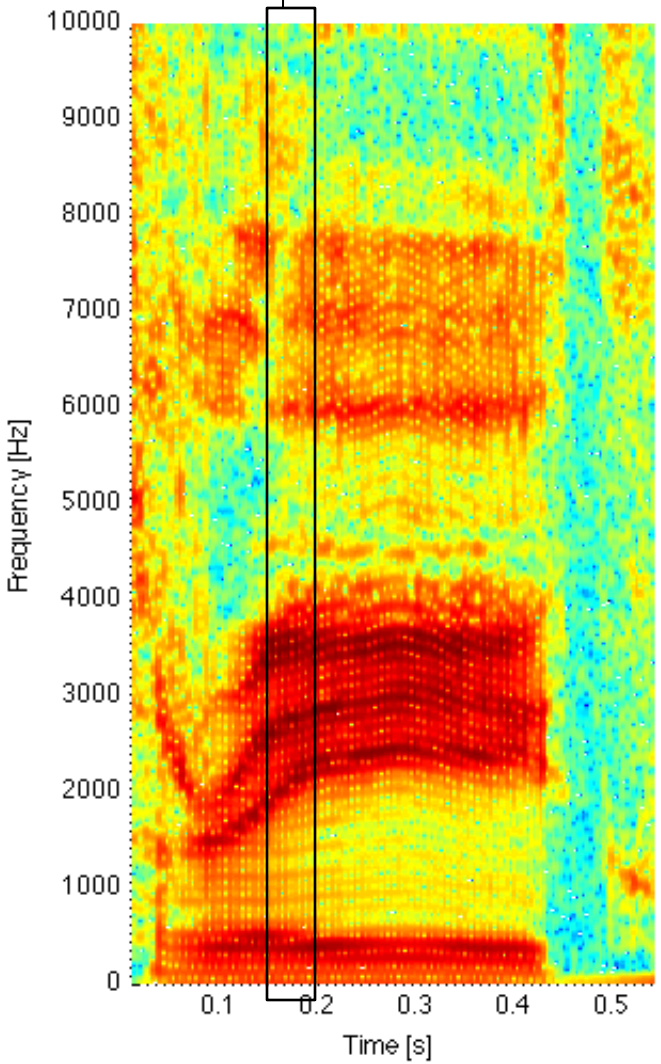


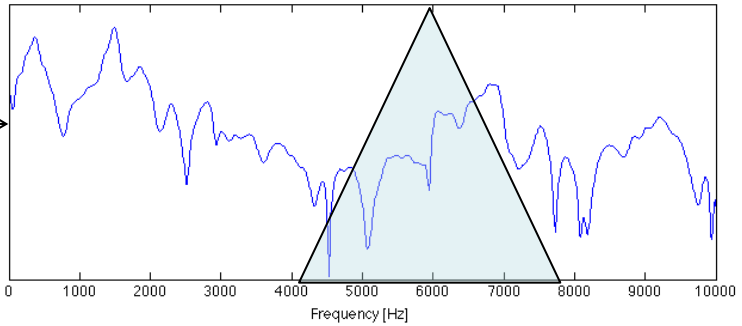
Forvinnsla


10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency [Hz]
0.1
0.2
0.3
0.4
0.5
Time [s]

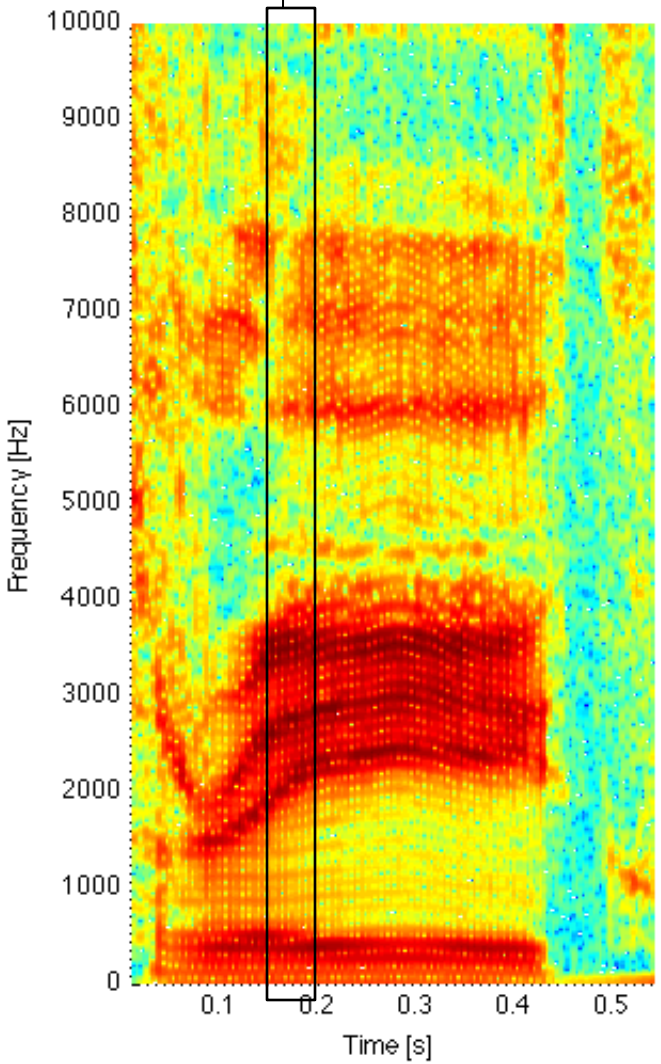

0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Frequency [Hz]

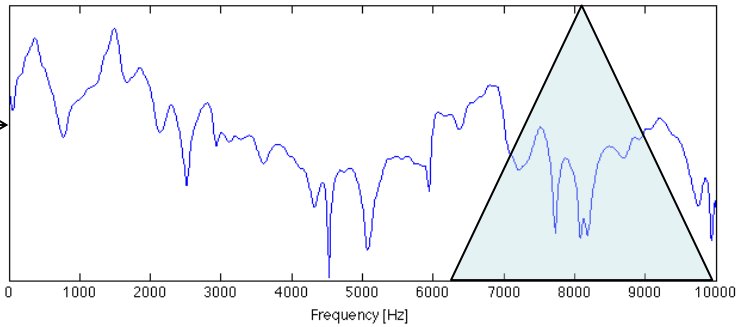
Forvinnsla

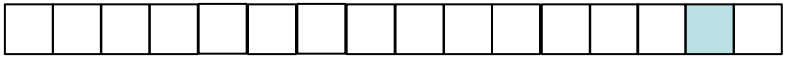

10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency [Hz]
0.1
0.2
0.3
0.4
0.5
Time [s]


0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Frequency [Hz]

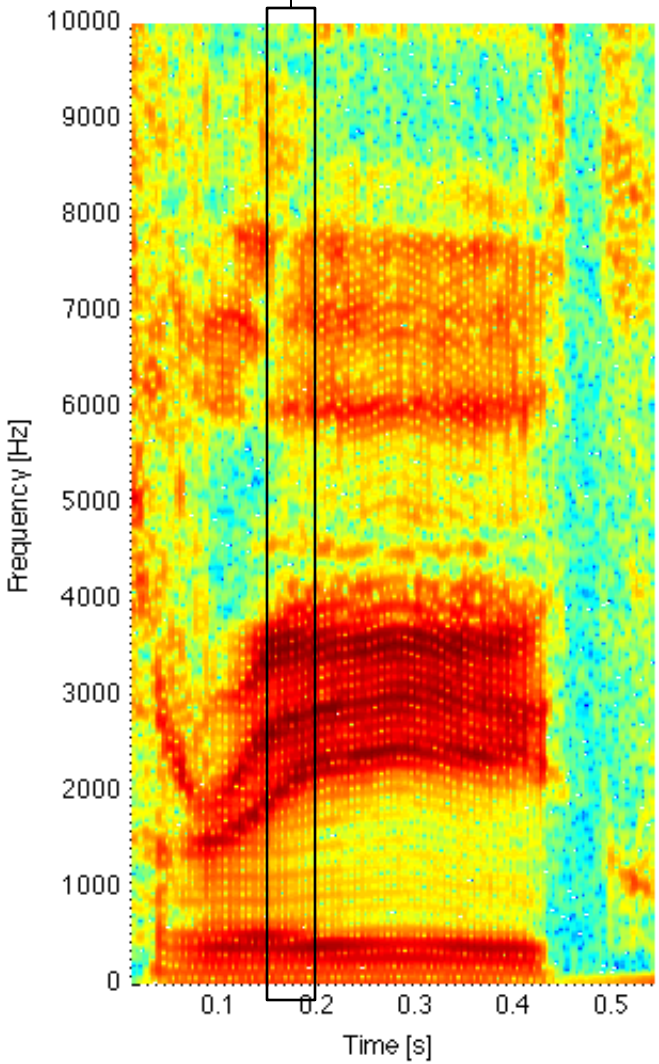
Forvinnsla

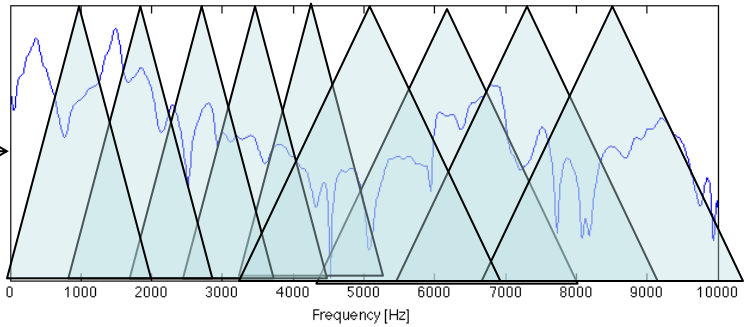

10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency [Hz]
0.1
0.2
0.3
0.4
0.5
Time [s]

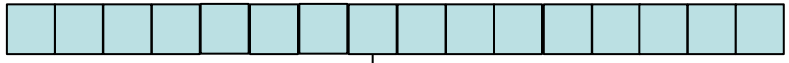

0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Frequency [Hz]



Forvinnsla

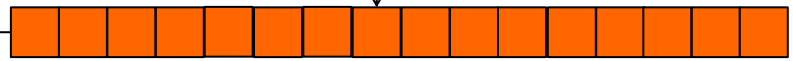




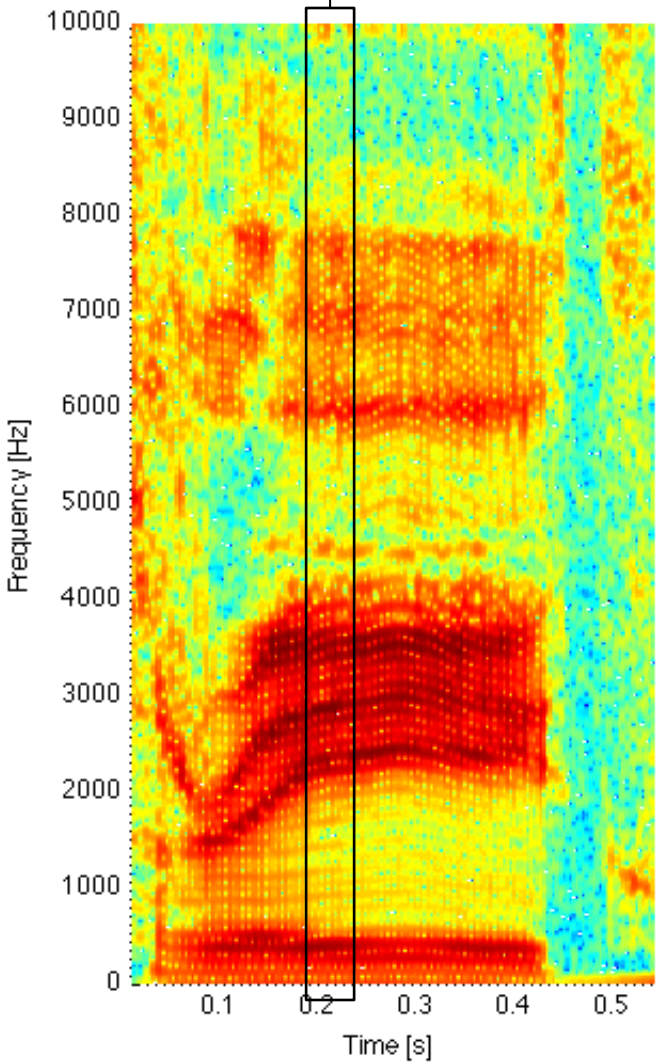


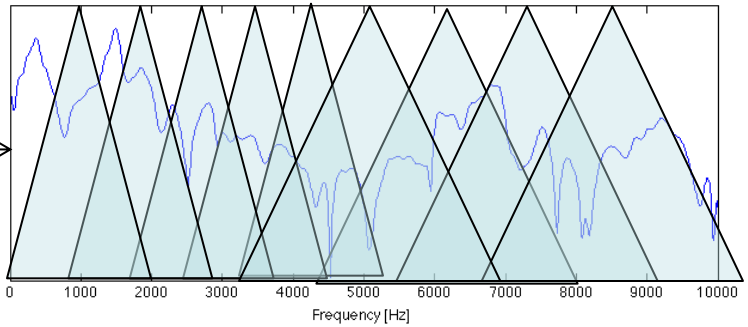
Kósínus vörpun

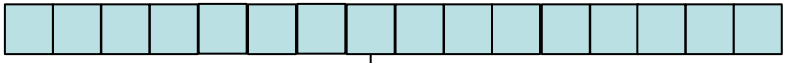
Melcepstrum



Forvinnsla

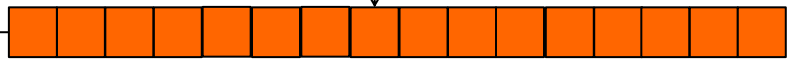


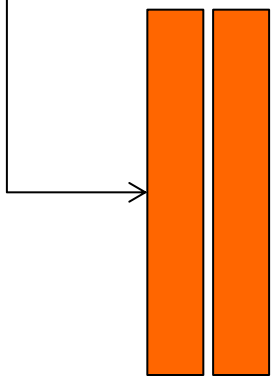




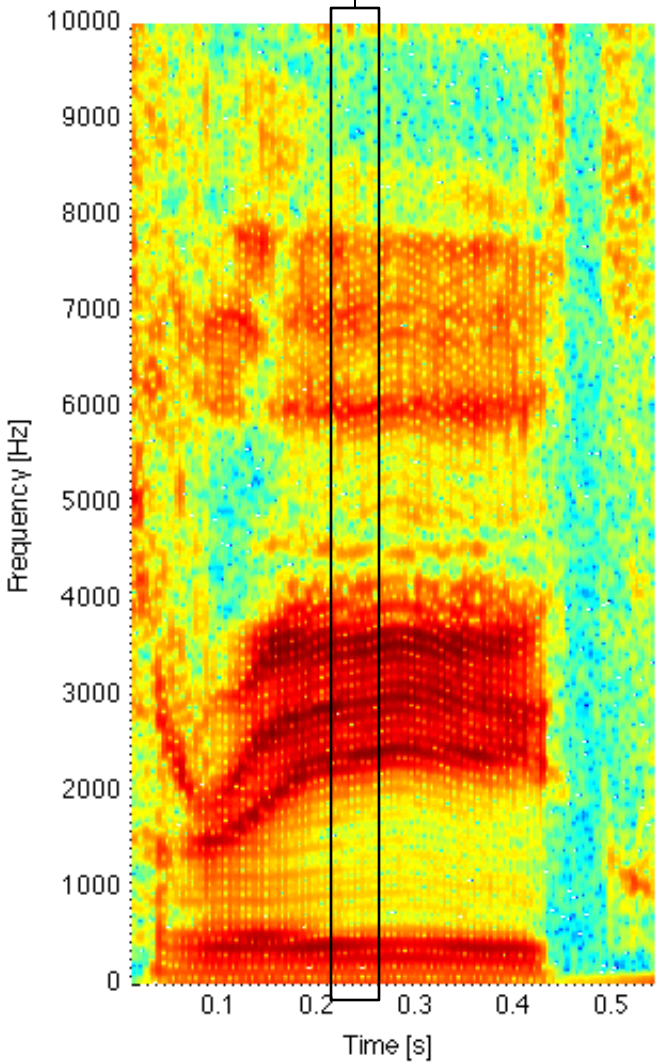
Kósínus vörpun

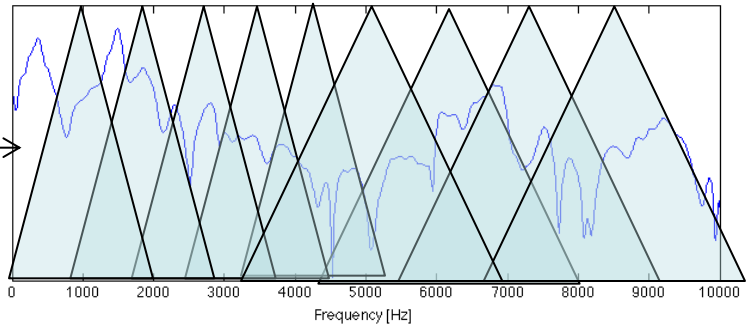
Melcepstrum

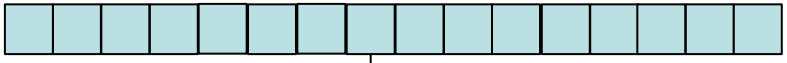




Forvinnsla

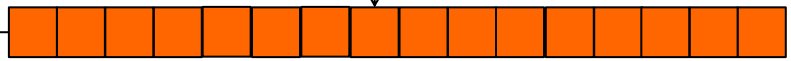


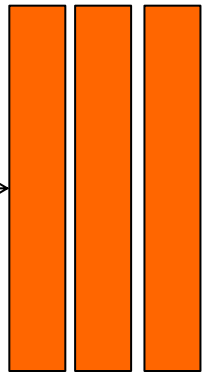




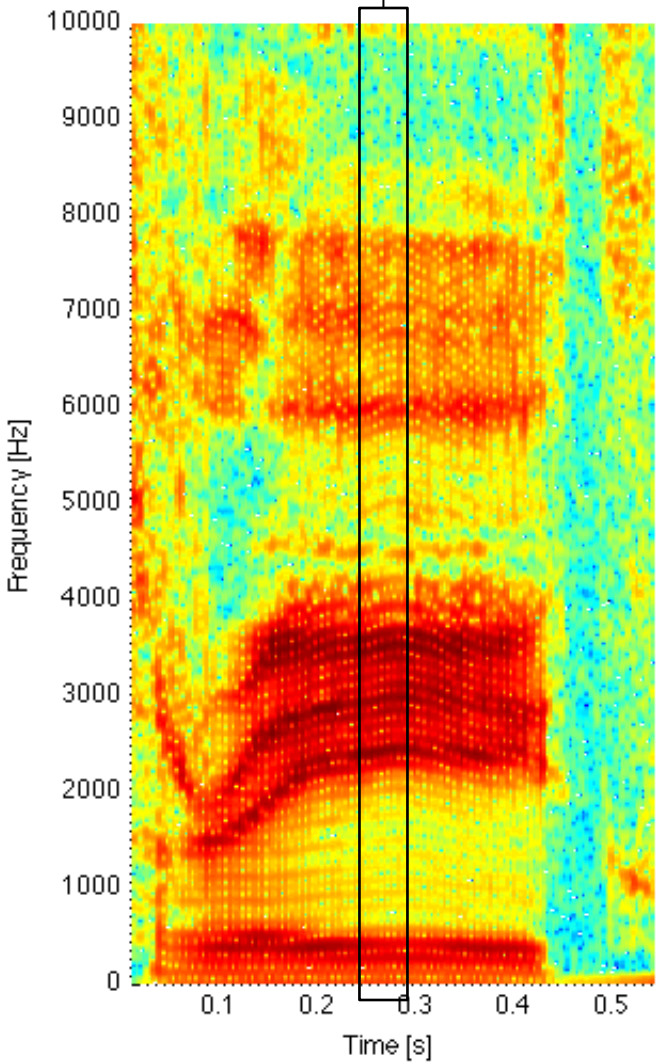
Kósínus vörpun

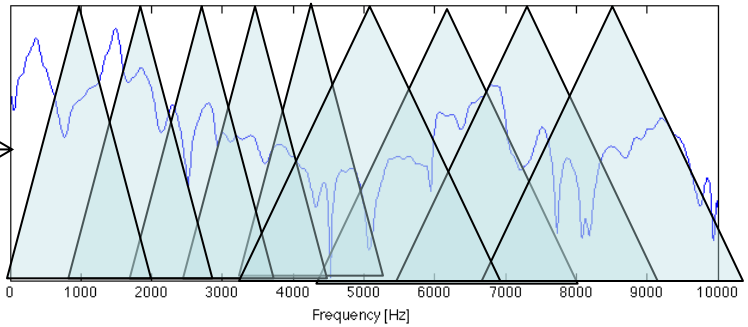
Melcepstrum

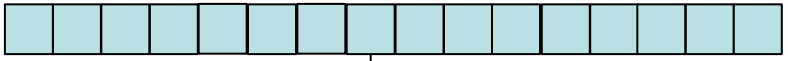




Forvinnsla

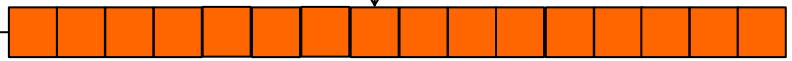


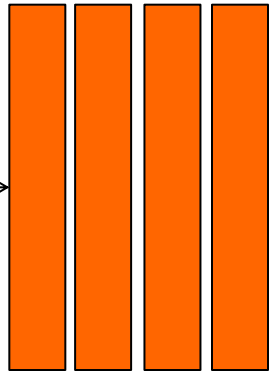




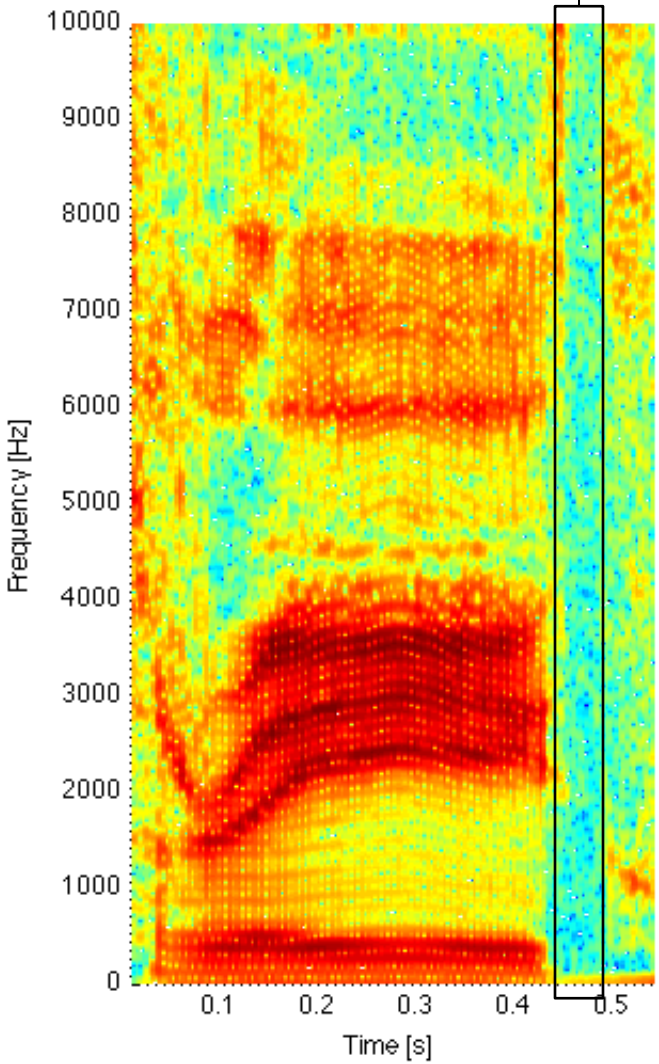
Kósínus vörpun

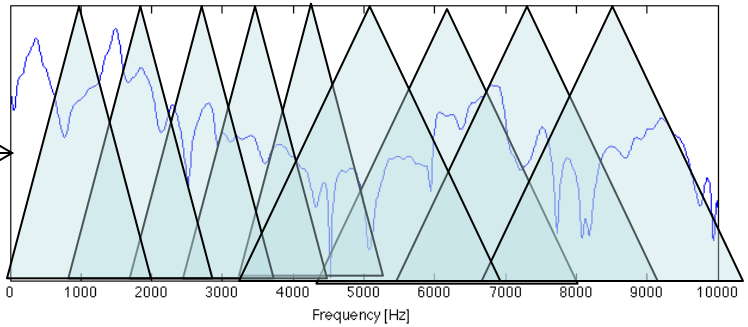
Melcepstrum

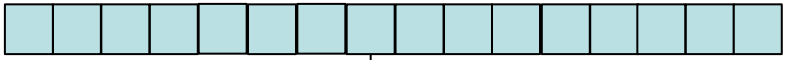




Forvinnsla

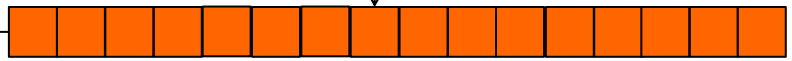




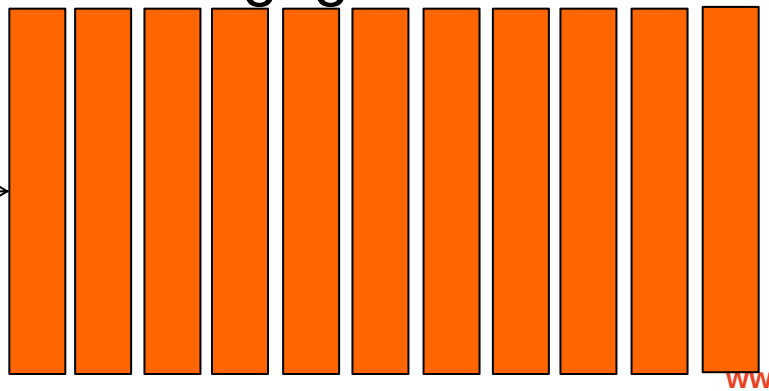


Kósínus vörpun

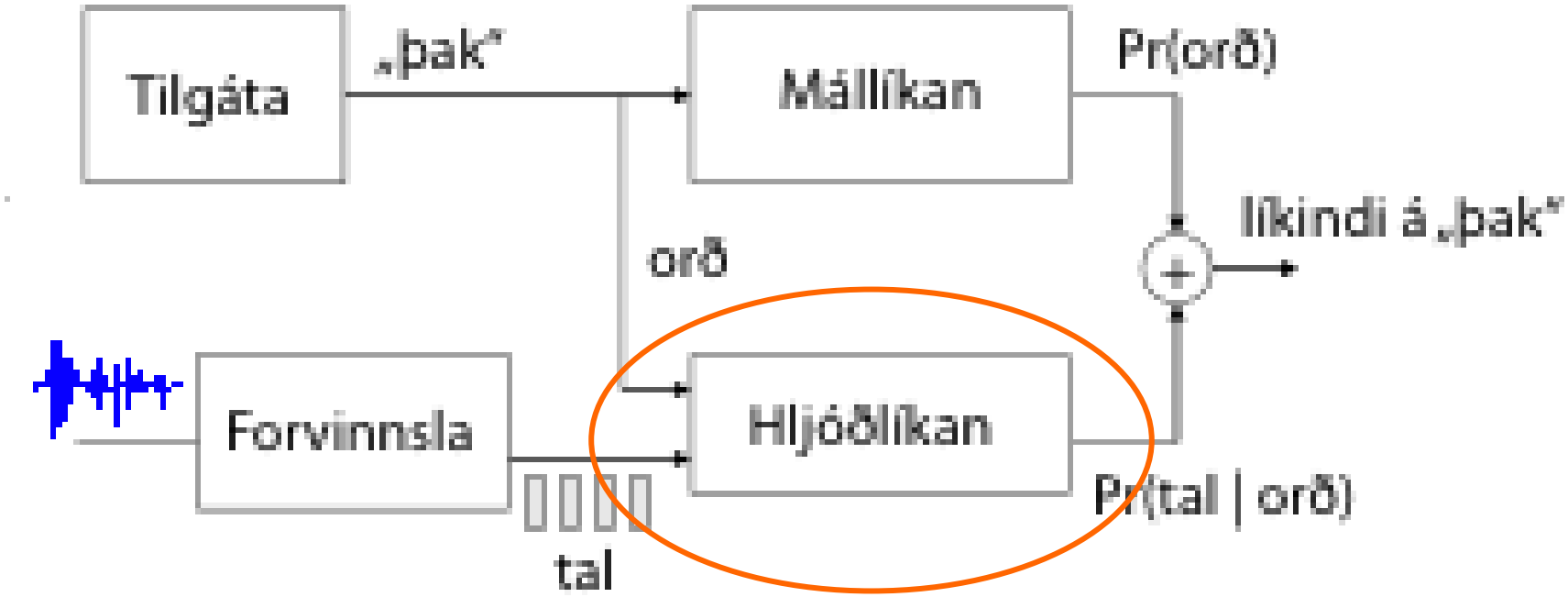
Melcepstrum



Runa af gagnarömmum

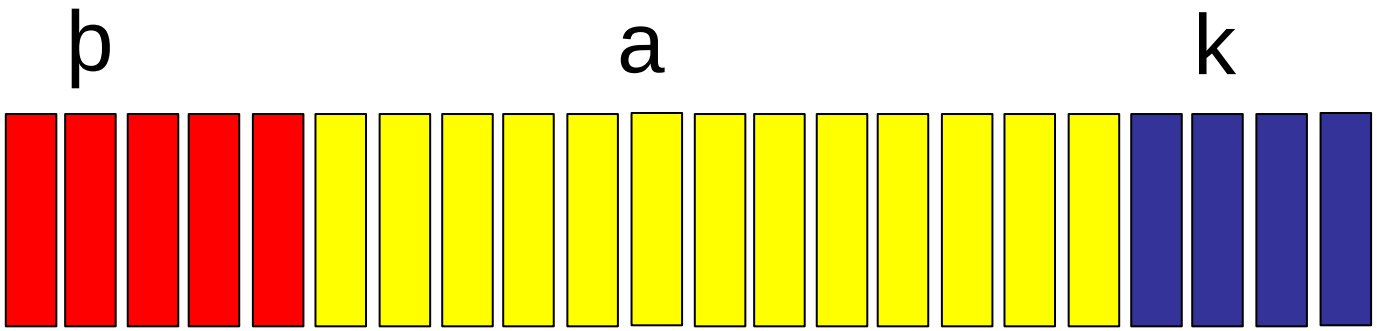


Talgreining og gagnaöflun

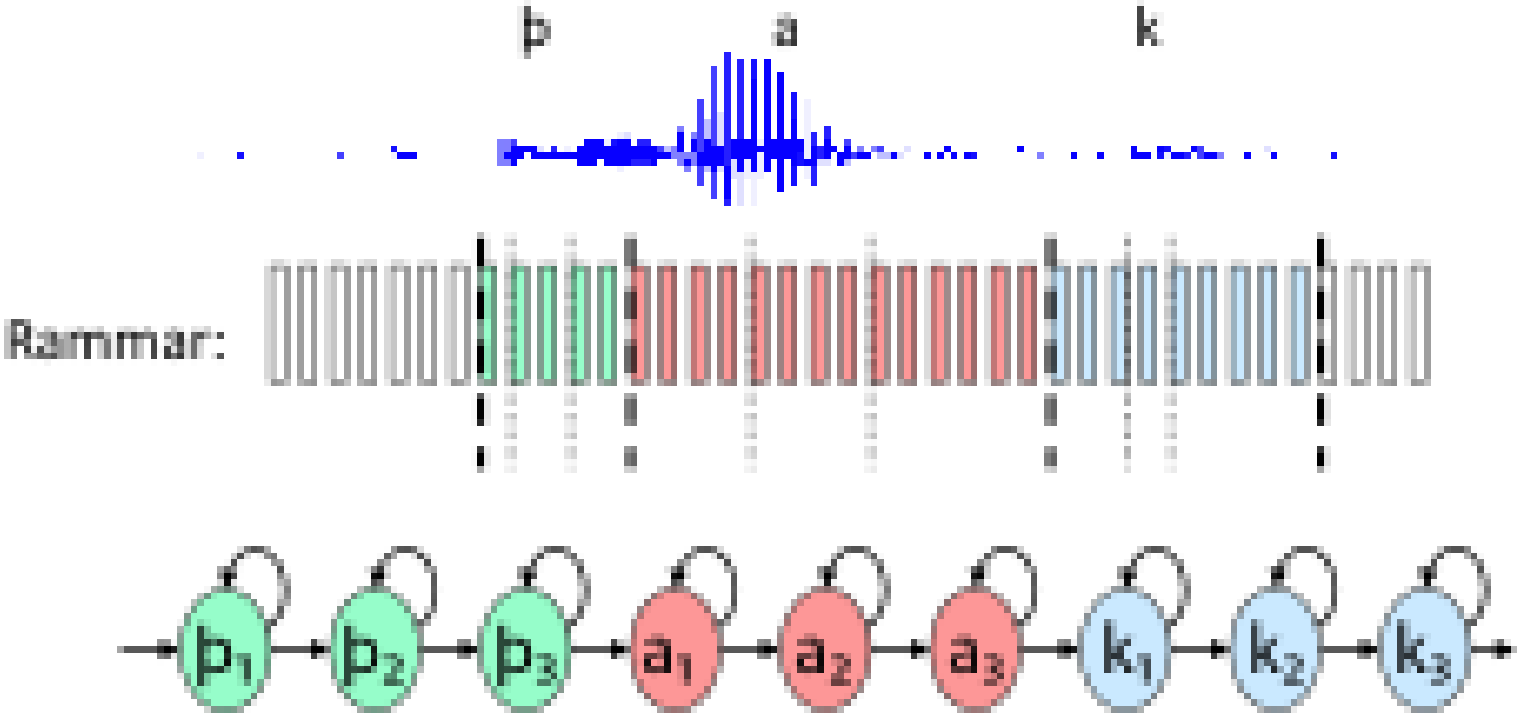


Rammagreining

- Greinum hvern ramma í gagnarununni
- Tölfræðileg líkön eða tauganet
  - Þarf mikið af gögnum til þess að þjálfa!



Greining rununnar



Við vitum hvað þarf til þess að þróa talgreini

- Gögn – þrír gagnagrunnar. Getum safnað meiru.
- Þróunarumhverfi – Þjálfun hægt að framkvæma með opnum hugbúnaði.
- Þekkingu – Þróum okkar eigin hugbúnað: Eign á hugverki
- Sérfræðingar og tími – Mannmánuðir í verkefni

Tækni sem nýtist samfélaginu og atvinnulífinu

- Nákvæmni
  - ásættanleg orðavillutíðni > 80%
- Rauntímakröfur
  - kerfið þarf að geta skilað niðurstöðum hratt og örugglega
- Miðlægur hugbúnaður?
  - Almannarómur þarf að reka miðlæga þjónustu fyrir almenning
- Þjónusta við þá sem þróa endahugbúnað
  - Fyrirtæki á Íslandi eiga að geta þróað hugbúnað byggðan á talgreiningu (og máltækni almennt)

Spurningar?