Linguistic Calculations on Cyfronet High Performance Computers

Bartosz Ziółko, Jakub Gałka,
Mariusz Ziółko

Katedra Elektroniki

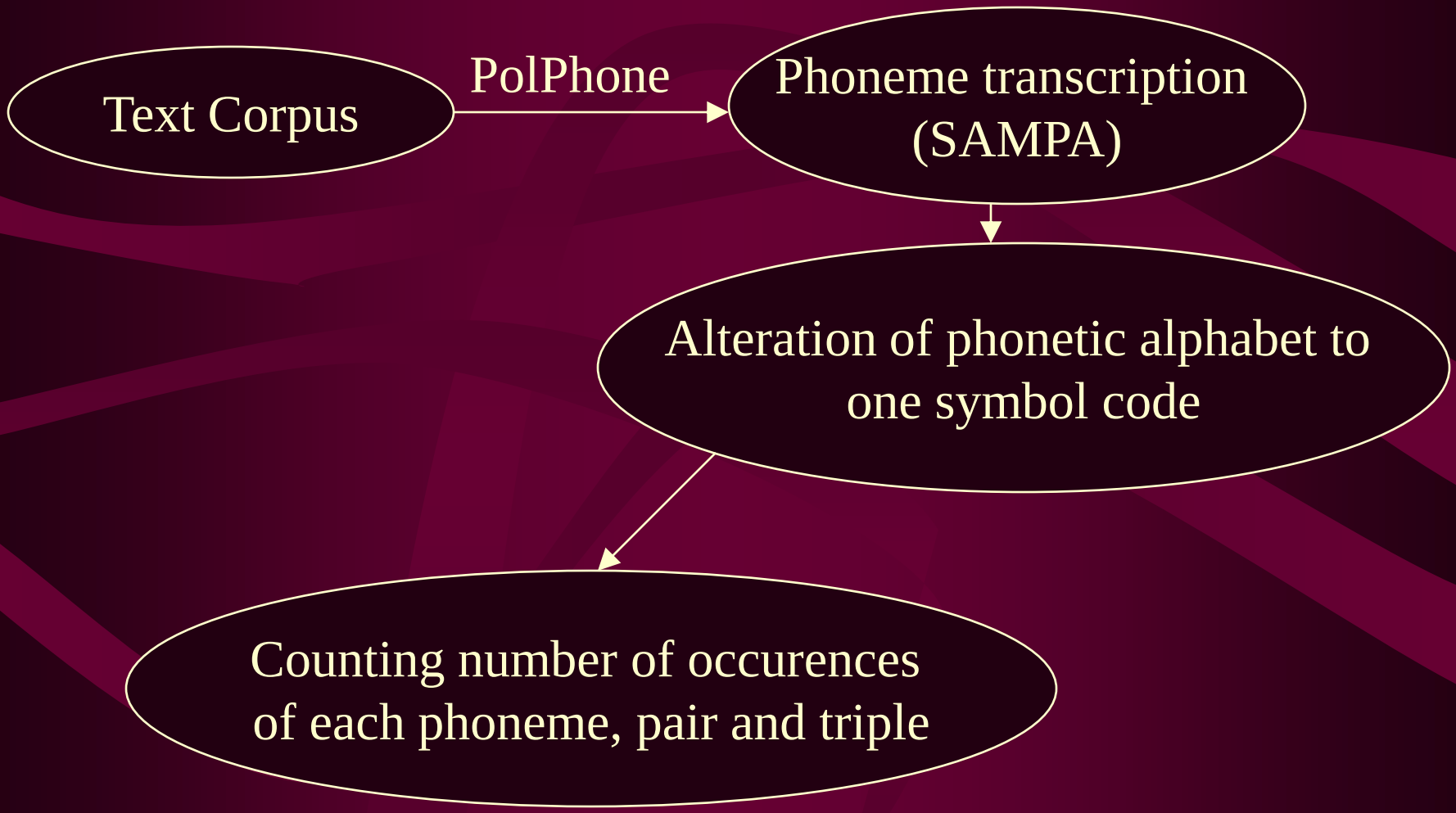
Zespół Przetwarzania Sygnałów

www.dsp.agh.edu.pl

AGH

Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Scheme of the experiment



SAMPA and phoneme frequencies

SAMPA	example	transcr.	occurr.	%
#		#	67,909,570	16.28
e	test	test	34,933,284	8.37
a	pat	pat	33,819,855	8.10
o	pot	pot	31,743,727	7.61
j	jak	jak	14,683,820	3.52
I	typ	tIp	14,367,038	3.44
t	test	test	13,980,824	3.35
i	PIT	pit	13,833,809	3.31
n	nasz	naS	13,749,670	3.29
m	mysz	mIS	12,179,292	2.91
v	wilk	vilk	11,777,111	2.82
r	ryk	rIk	11,696,445	2.80
p	pik	pik	11,281,812	2.70
u	puk	puk	10,578,340	2.53
w	łyk	wIk	10,104,187	2.42
s	syk	sIk	9,793,251	2.34
d	dym	dIm	9,140,704	2.19
n’	koń	kon’	8,547,530	2.05
k	kit	kit	8,435,010	2.02
l	luk	luk	7,844,660	1.88
z	zbir	zbir	7,136,927	1.71
g	gen	gen	5,984,361	1.43
b	bit	bit	5,897,286	1.41
S	szyk	SIk	5,870,091	1.41
s’	świt	s‘vit	5,391,461	1.29
Z	żyto	ZIto	4,827,820	1.16
f	fan	fan	4,596,380	1.10
ts	cyk	tsIk	4,002,641	0.96
x	hymn	xImn	3,944,391	0.95
ts’	ćma	ts’ma	3,845,071	0.92
tS	czyn	tSIn	3,731,910	0.89
dz’	dźwig	dz’vik	3,235,969	0.78
w∼	ciąża	ts’ow∼Za	2,579,732	0.62
c	kiedy	cjedy	1,962,446	0.47
dz	dzwoń	dzvon’	1,028,028	0.25
z’	źle	z’le	996,629	0.24
N	pęk	peNk	833,599	0.20
c	kiedy	cjedy	507,679	0.12
dZ	dżem	dZem	201,248	0.05
j∼	więź	vjej∼s’	154,452	0.04

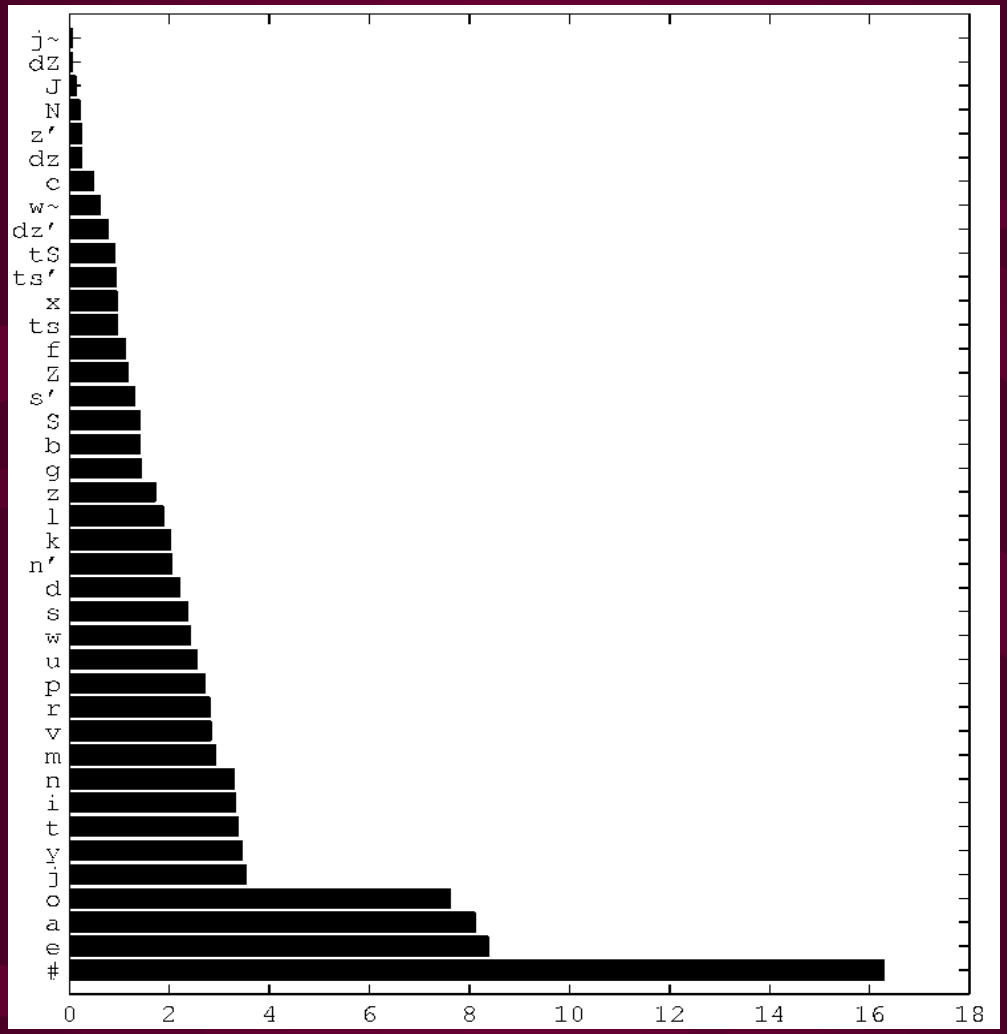
Cyfronet resources we used

Mars is a high performance computer with following specication: IBM Blade Center HS21 - 112 Intel Dual-core processors, 8GB RAM/core, 5 TB disk storage and 1192 Gops. It operates using Red Hat Linux.

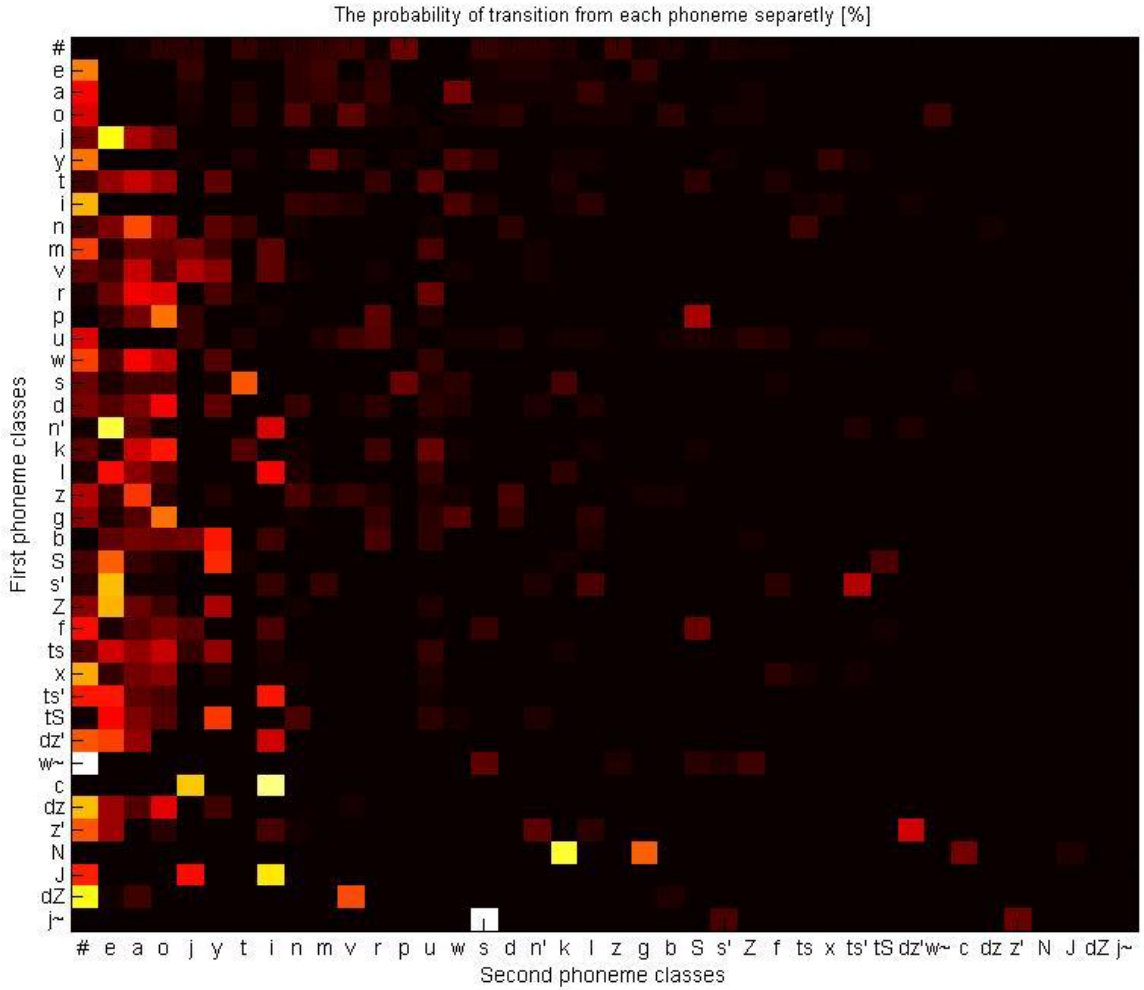
Saturn uses Solaris 10 and has specication: Sun Fire 6800 12 UltraSparc III processors (900 MHz clock), 12 GB RAM, 120 GB + 1 TB disk storage and 36 Gops.

All calculations were conducted in Matlab.

Phoneme statistics



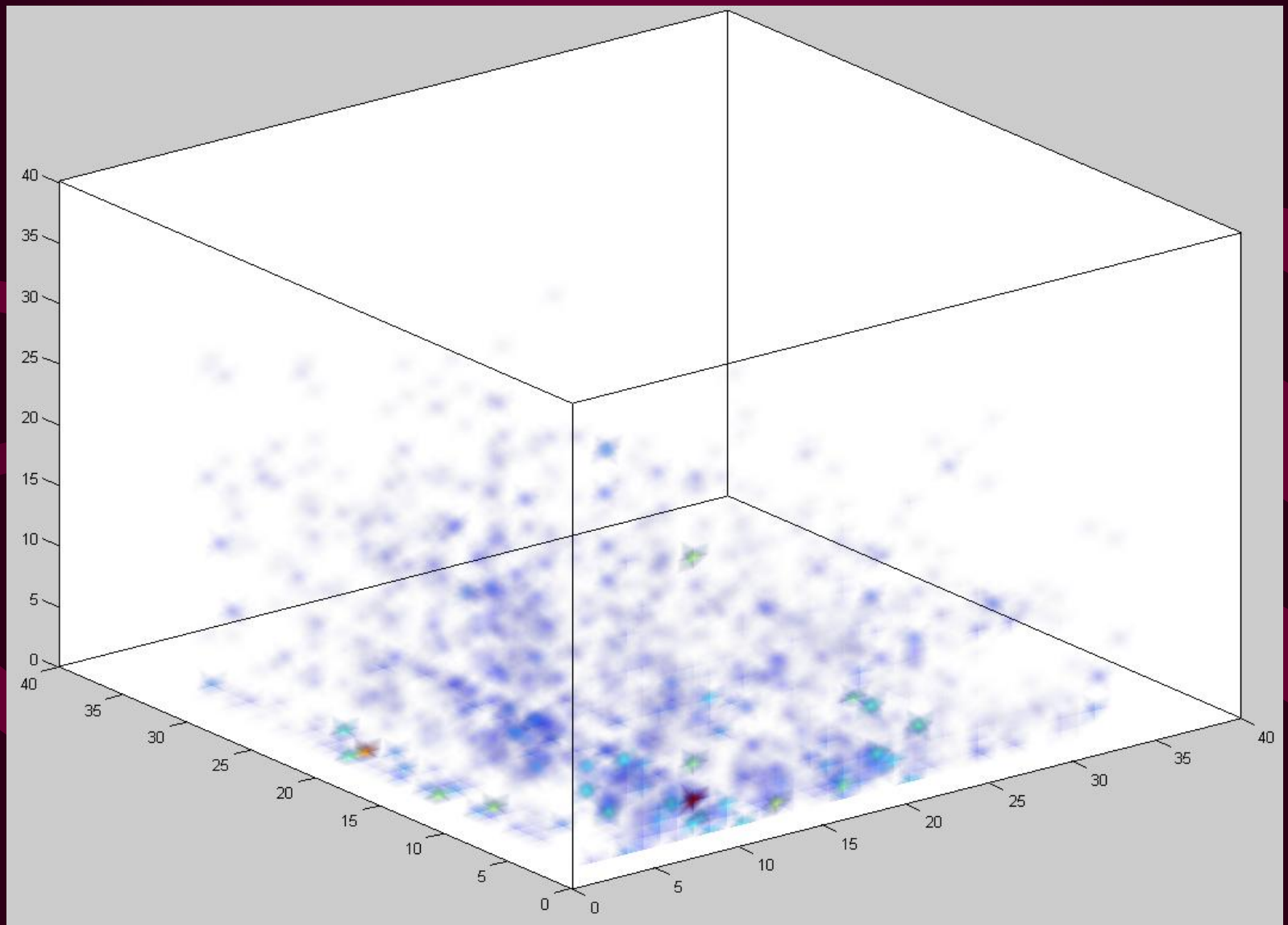
Diphones


The probability of transition from each phoneme separetly [%]
First phoneme classes
e a o j y t i n m v r p u w s d n' k l z g b S s' Z f ts x ts' tS dz' w~ c dz z' N J dZ j~
Second phoneme classes

Most common diphones

diphone	no. of occurrences	percentage
e#	12,652,597	3.034
a#	8,141,149	1.952
#p	7,369,012	1.767
je	7,326,862	1.757
o#	6,887,824	1.652
i#	5,704,800	1.368
y#	5,124,797	1.229
n'e	4,525,089	1.085
#z	4,404,026	1.056
na	4,314,733	1.035
#v	4,293,464	1.029
#t	4,028,657	0.966
po	4,028,172	0.966
#s	3,973,928	0.953
aw	3,731,959	0.895
m#	3,670,595	0.880
#m	3,670,134	0.880
st	3,138,007	0.752
#o	3,109,260	0.745
w#	3,104,722	0.744
#d	3,010,451	0.722
#j	3,005,071	0.720
ov	2,938,270	0.704
#n	2,896,448	0.694
#n'	2,845,016	0.682
on	2,777,159	0.666
ra	2,711,110	0.650
ta	2,686,110	0.644
#s'	2,600,191	0.623
ro	2,557,600	0.613
ja	2,491,371	0.597
wa	2,457,503	0.589
#b	2,431,739	0.583
#k	2,412,680	0.578
em	2,377,256	0.570
#i	2,334,027	0.560
va	2,326,907	0.558
s'e	2,267,362	0.544
do	2,264,599	0.543
u#	2,228,523	0.534
ko	2,228,041	0.534
ow~	2,126,896	0.510
go	2,121,696	0.509

Triphones



Most common triphones

#po	3,171,836	0.761
n'e#	3,017,727	0.724
#na	2,537,946	0.609
#n'e	2,205,296	0.529
#s'e	2,038,733	0.489
na#	2,017,989	0.484
s'e#	1,952,149	0.468
#za	1,867,754	0.448
ow~#	1,841,737	0.442
#pS	1,672,570	0.401
vje	1,662,129	0.399
#i#	1,639,503	0.393
go#	1,637,610	0.393
#do	1,629,173	0.391
#je	1,617,416	0.388
em#	1,604,536	0.385
aw#	1,564,615	0.375
je#	1,498,358	0.359
wa#	1,468,262	0.352
ej#	1,422,285	0.341
ego	1,406,321	0.337
e#p	1,394,315	0.334
Ze#	1,229,760	0.295
#vy	1,162,213	0.279
pSe	1,093,322	0.262
#Ze	1,062,468	0.255
ova	1,019,807	0.245
sta	1,013,048	0.243
e#z	986,469	0.237
#to	967,113	0.232
#ja	963,055	0.231
to#	956,517	0.229
ym#	923,397	0.221
a#p	903,202	0.217
e#v	895,543	0.215
#st	879,684	0.211
li#	861,925	0.207
mje	861,119	0.206
#by	853,189	0.205
cje	848,842	0.204
awa	848,323	0.203
le#	834,275	0.200
do#	831,737	0.199
e#m	820,973	0.197
#te	816,063	0.196
#f#	789,214	0.189
jon	786,661	0.189
#v#	780,416	0.187
#pa	775,681	0.186
#ta	772,413	0.185
e#s	766,573	0.184
#mo	752,955	0.181
ne#	737,168	0.177
o#p	736,683	0.177

Input Data

- Transcriptions of Parliment and committees meetings,
- Literature
- Internet articles
- Newspaper articles

In the first experiment a total number of 148,016,538 phonemes were analysed, what took 3 weeks using Matlab on PC

Now we have corpora containg around 2,000,000,000

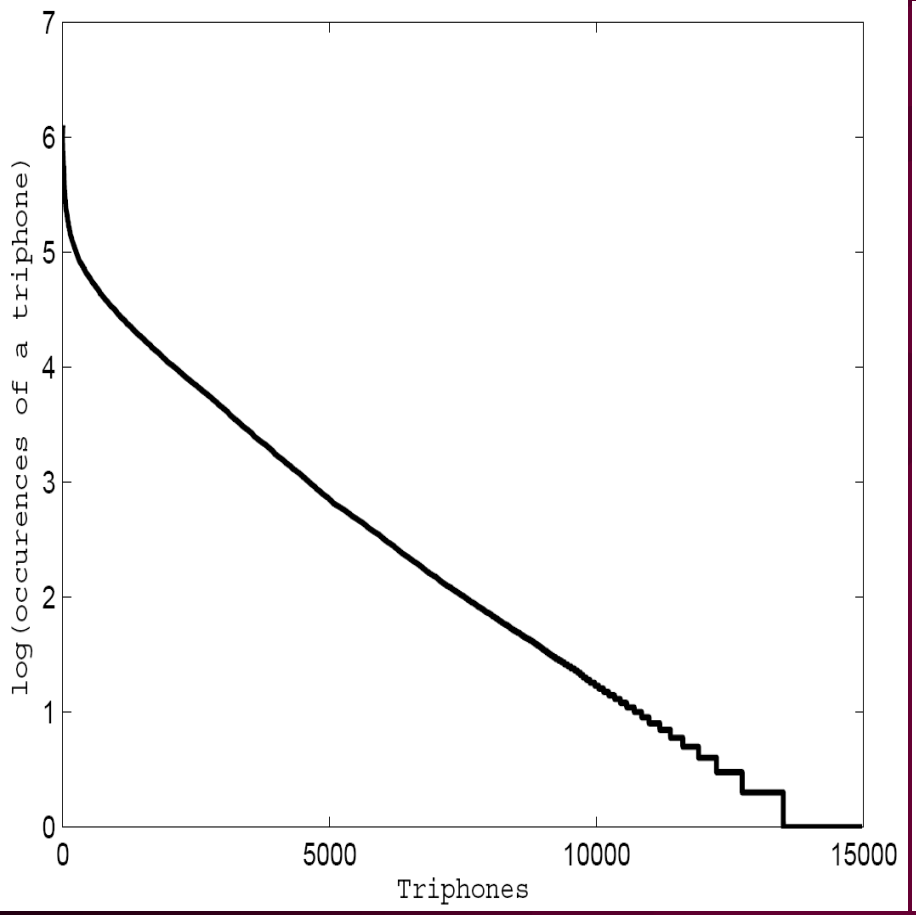
And still collect more ... (bziolko@agh.edu.pl)

Results and some observations

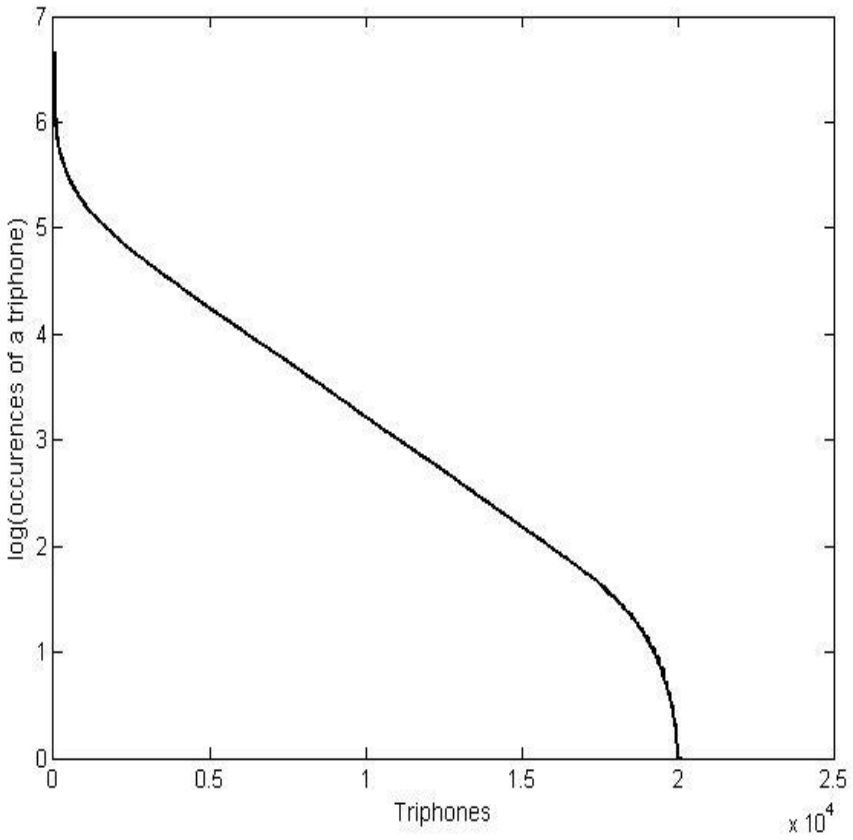
- Around 1,250 different diphones were detected
- Over 20,000 different triphones out of 62,479 possible combinations (excluding phoneme space phoneme string) were detected (32%)

Distribution of frequencies

PC



Mars



Conclusions

- Triphone statistics play an important role in speech recognition systems,
- 32% of possible triphones were detected,
- Some of them were very rare and came from foreign and twisted words,
- Usage of High Performence Computers is necessary for modern tasks in speech and natural language processing.

Thank You