Speech Recognition of Highly Inflective Languages

Bartosz Ziółko,
University of York

Research structure

- Literature review
- Linguistic aspects of Polish
  - HTK for Polish, triphone statistics
- Phoneme segmentation and acoustic models
  - Method based on DWT, fuzzy recall and precision, SVM and LogitBoost Weka classifier
- Language models
  - POS tagging and semantic model

ASR scheme

Noise reduction

Segmentation

Parametrisation

Trained HMMs

Phoneme recogniser

Word level network

Dictionary

Sentence level network

Syntax model

Semantic coherence

LSA model

ASR of Polish

- Much fewer homophones
- More distinguishable sounds of vowels
- Very few irregularities in pronunciation
- Rustling phonemes
- Complicated grammar and morphology
- Inflective, not positional
- Possibly smaller dictionary

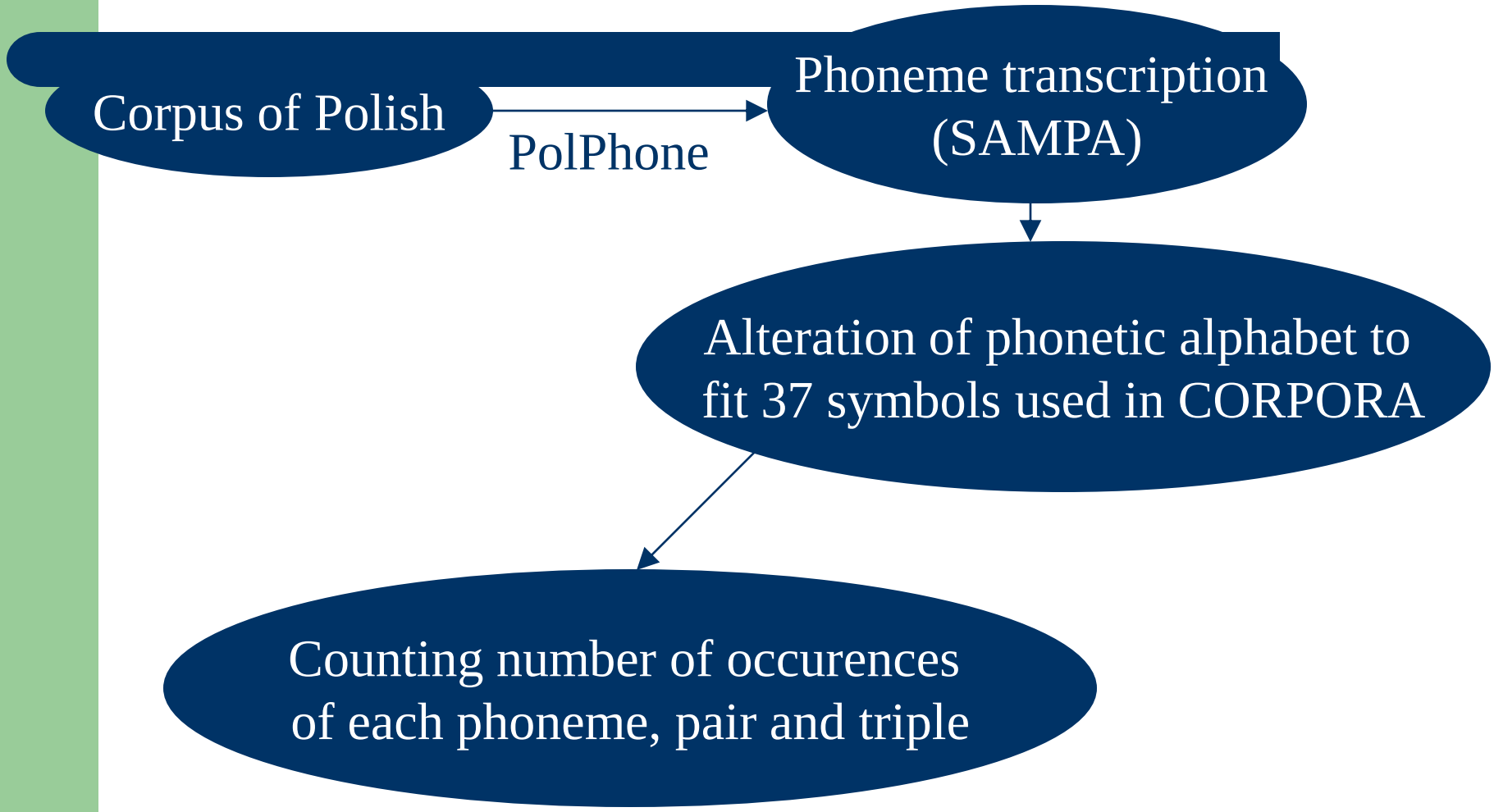
HMM Toolkit (HTK)

- Trained on 26 male adult speakers (9490 utterances)
- frequency 16 kHz
- 39 MFCCs
- 25 ms windows
- preempasis filtering 0.97

Corectness for different speakers

speaker	age	gender	substitutions	correctness
AO1M1	adult	male	6	98.36
AF1K1	adult	female	74	79.73
BC1K1	adult	female	17	95.34
BW1K1	adult	female	29	92.05
AK1C1	child	male	144	60.55
AK2C1	child	male	89	75.62
CK1C1	child	male	18	95.07
LK1D1	child	female	43	88.22
ZK1D1	child	female	58	84.11

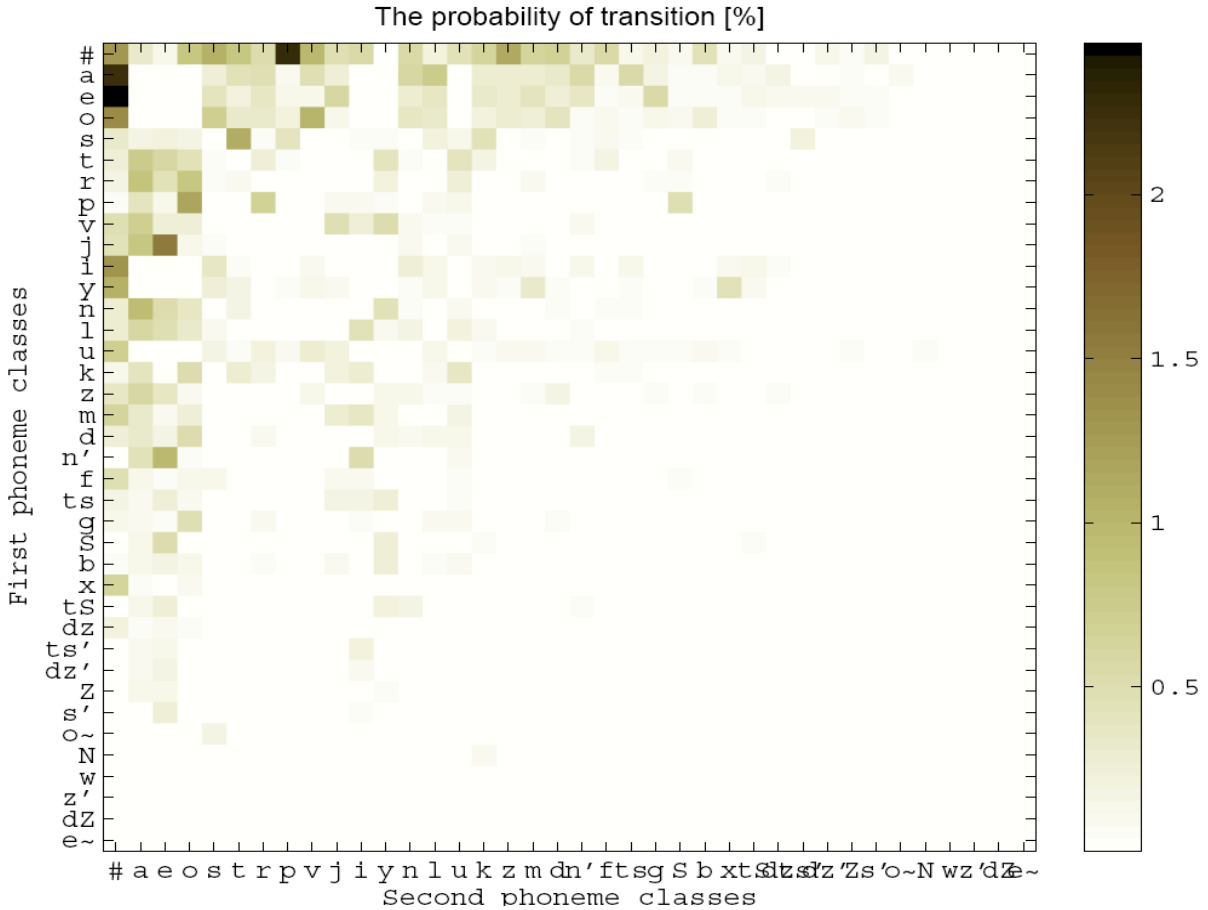
Triphone statistics



Sampa and phoneme frequencies

SAMPA	example	transcr.	occurr.	%
#		#	23,810,956	16.086,7
a	pat	pat	13,311,163	8.993
e	test	test	11,871,405	8.020,3
o	pot	pot	10,566,010	7.138,4
s	syk	sIk	5,716,058	3.861,8
t	test	test	5,703,429	3.853,2
r	ryk	rIk	5,171,698	3.494
p	pik	pik	5,150,964	3.48
v	wilk	vilk	5,025,050	3.394,9
j	jak	jak	4,996,475	3.375,6
i	PIT	pit	4,994,743	3.374,4
I	typ	tIp	4,974,567	3.360,8
n	nasz	naS	4,602,314	3.109,3
l	luk	luk	4,399,366	2.972,2
u	puk	puk	4,355,825	2.942,8
k	kit	kitk	4,020,161	2.716
z	zbir	zbir	3,602,857	2.434,1
m	mysz	mIS	3,525,813	2.382
d	dym	dIm	3,267,009	2.207,2
n’	koń	kon’	3,182,940	2.150,4
f	fan	fan	2,030,717	1.372
ts	cyk	tsIk	1,984,311	1.340,6
g	gen	gen	1,949,890	1.317,3
S	szyk	SIk	1,739,146	1.175
b	bit	bit	1,668,103	1.127
x	hymn	xImn	1,339,311	0.904,84
tS	czyn	tSIn	1,285,310	0.868,36
dz	dzwoń	dzvon’	692,334	0.467,74
ts’	ćma	ts’ma	690,294	0.466,36
dz’	dźwig	dz’vik	589,266	0.398,11
Z	żyto	ZIto	536,786	0.362,65
s’	świt	s‘vit	531,402	0.359,02
o~	wąs	vo~s	306,665	0.207,18
N	pęk	peNk	184,884	0.124,91
w	łyk	wIk	144,166	0.097,399
z’	źle	z’le	66,518	0.044,94
dZ	dżem	dZem	27,621	0.018,661
e~	gęś	ge~s’	1,011	0.000,683
w~	ciąża	ts’ow~Za	sampa extension	
j~	więź	vjej~s’	sampa extension	
c	kiedy	cjedy	sampa extension	
J	giełda	Jjewda	sampa extension	

Diphones


The probability of transition [%]
First phoneme classes
Second phoneme classes
a e o s t r p v j i y n l u k z m d n' f ts g S b x tS dz ts' dz' Z s' o~ N w z' dZ e~
2
1.5
1
0.5

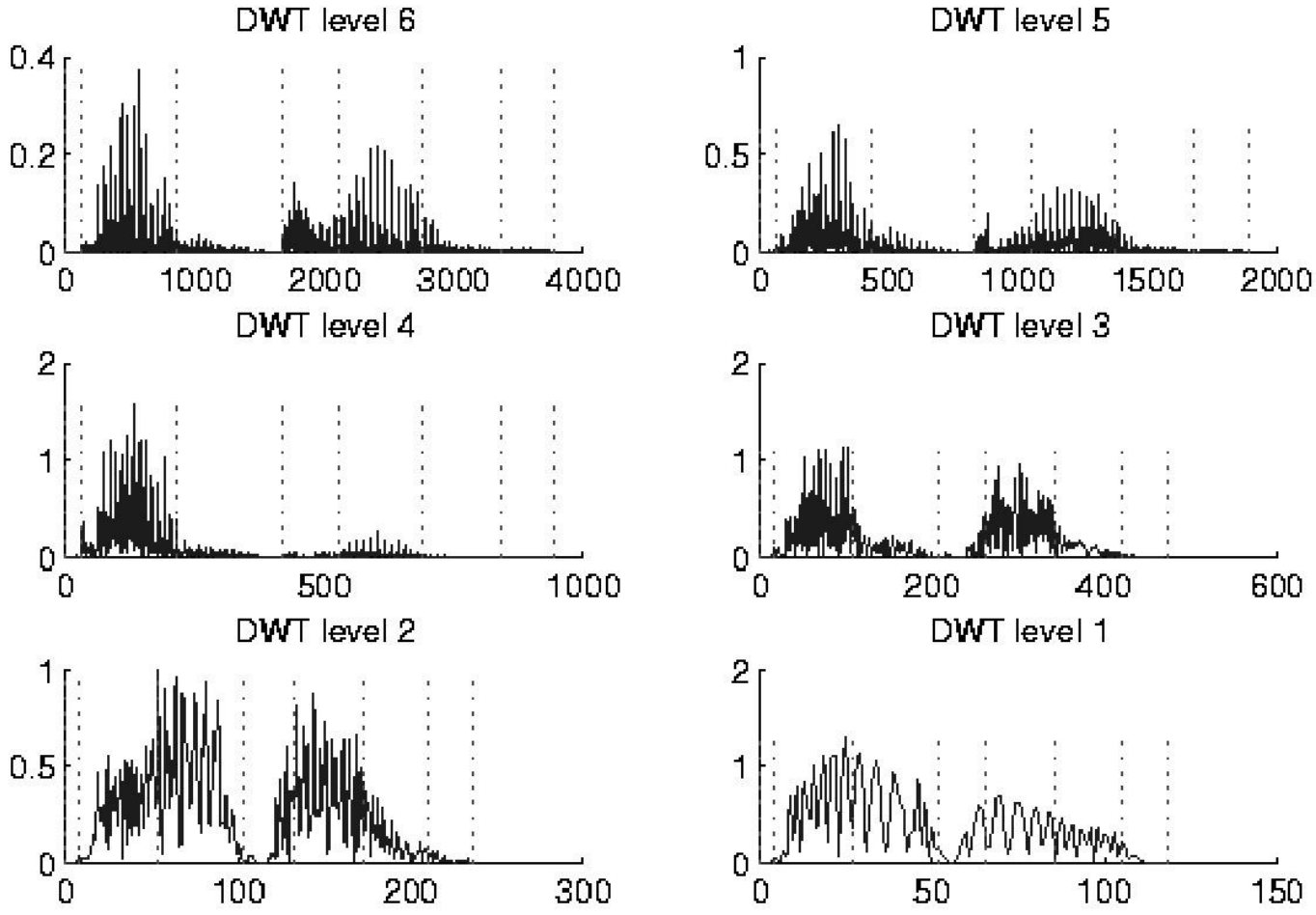
Most common triphones

triphone	no. of occurrences	percentage
#po	1,273,417	1.026,1
n'e#	925,893	0.746,09
#na	699,608	0.563,75
#pS	660,062	0.531,88
je#	659,674	0.531,57
na#	655,722	0.528,38
#pr	627,962	0.506,02
Ix#	613,589	0.494,43
ej#	602,920	0.485,84
#za	598,060	0.481,92
n'a#	574,708	0.46,31
ova	561,910	0.452,79
ego	558,788	0.450,27
sta	554,876	0.447,12
#do	551,423	0.444,34
go#	551,042	0.444,03
pSe	522,611	0.421,12
pra	492,128	0.396,56
#pa	481,772	0.388,21
#i#	478,500	0.385,58
vje	468,848	0.377,8
#n'e	430,178	0.346,64
#je	421,223	0.339,42
#f#	416,467	0.335,59
#v#	412,967	0.332,77
#vy	407,092	0.328,04
pro	390,429	0.314,61
#sp	357,008	0.287,68
#ko	342,254	0.275,79
#te	341,900	0.275,5
an'e	338,530	0.272,79
pos	337,190	0.271,71
ze#	335,941	0.270,7
ym#	332,437	0.267,88
em#	328,629	0.264,81
rav	318,232	0.256,43
#ze	310,008	0.249,81
ne#	309,151	0.249,12
nyx	307,657	0.247,91
kje	304,426	0.245,31
do#	296,635	0.239,03
ja#	294,220	0.237,08
#st	291,797	0.235,13
s'e#	285,355	0.229,94
#o#	283,500	0.228,45
ki#	282,413	0.227,57
#ro	282,059	0.227,28
to#	272,585	0.219,65
an'a	270,668	0.218,11
mje	266,812	0.215
ktu	265,128	0.213,64
#s'e	257,323	0.207,35

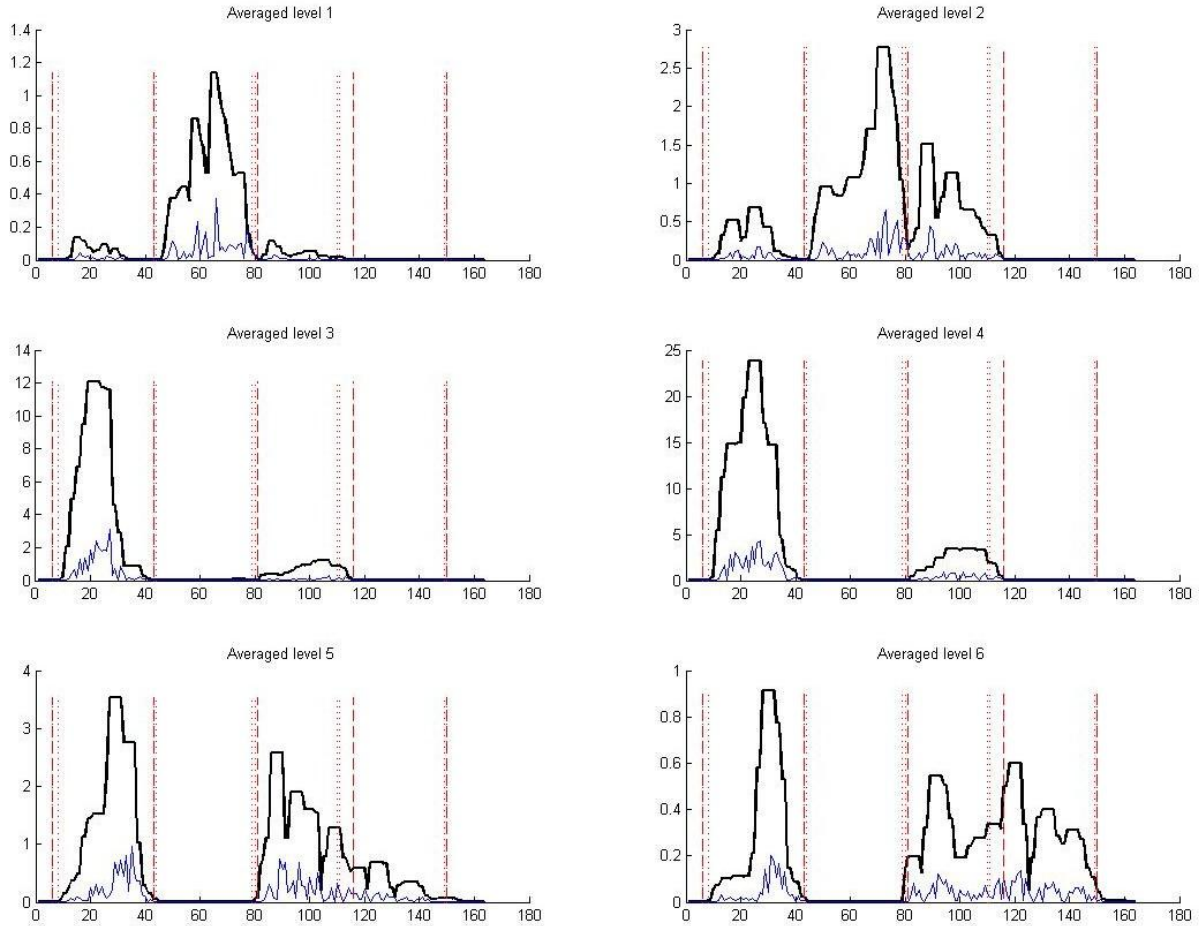
Results and some observations

- 1,095 different diphones were detected
- 14,970 different triphones out of 53,503 possible combinations (excluding phoneme space phoneme string) were detected (28%)
- Anomalia in corpus – The word „poseł” appeared 141,904 times in just its morphologically basic form which is 11% of total appearance of *#po* and 42% of *pos*
- Average length of words in phonemes is 6.2 (space frequency is 16.09 %)

DWT spectrum of speech signal



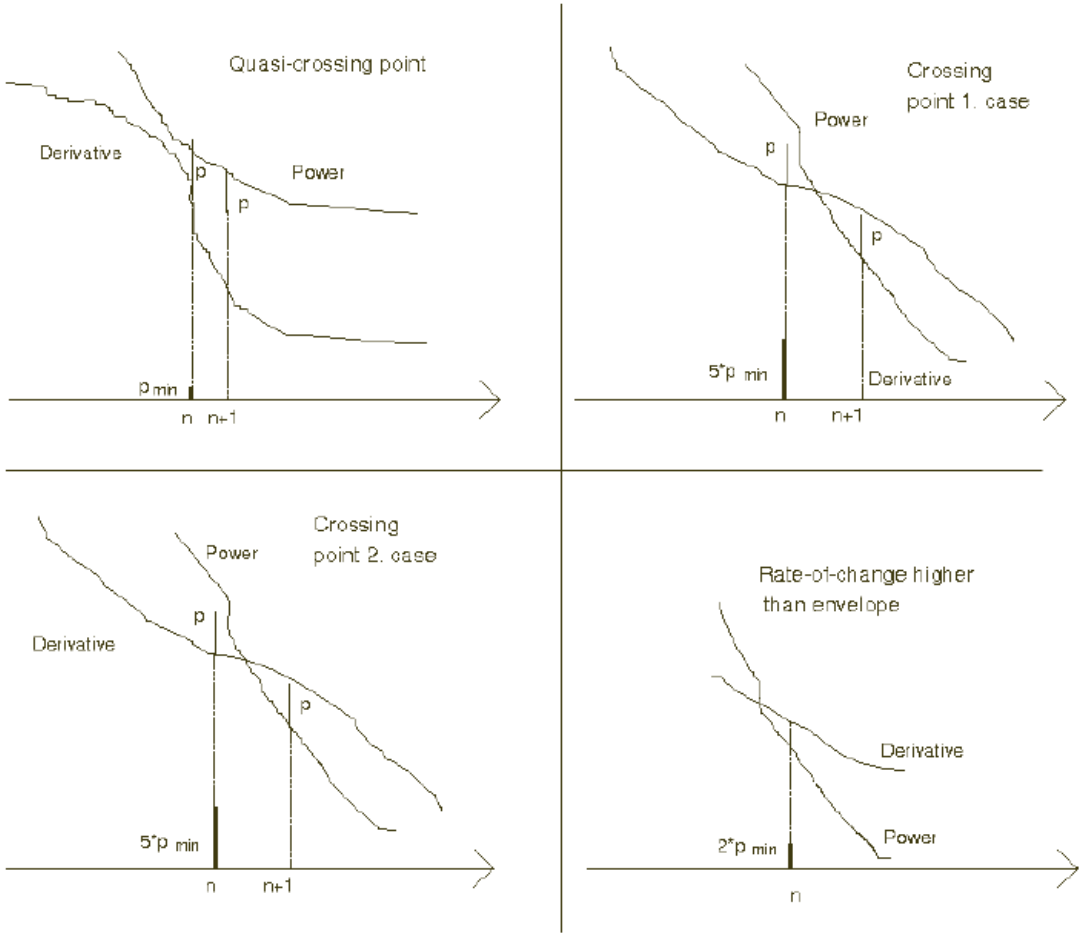
Phoneme segmentation



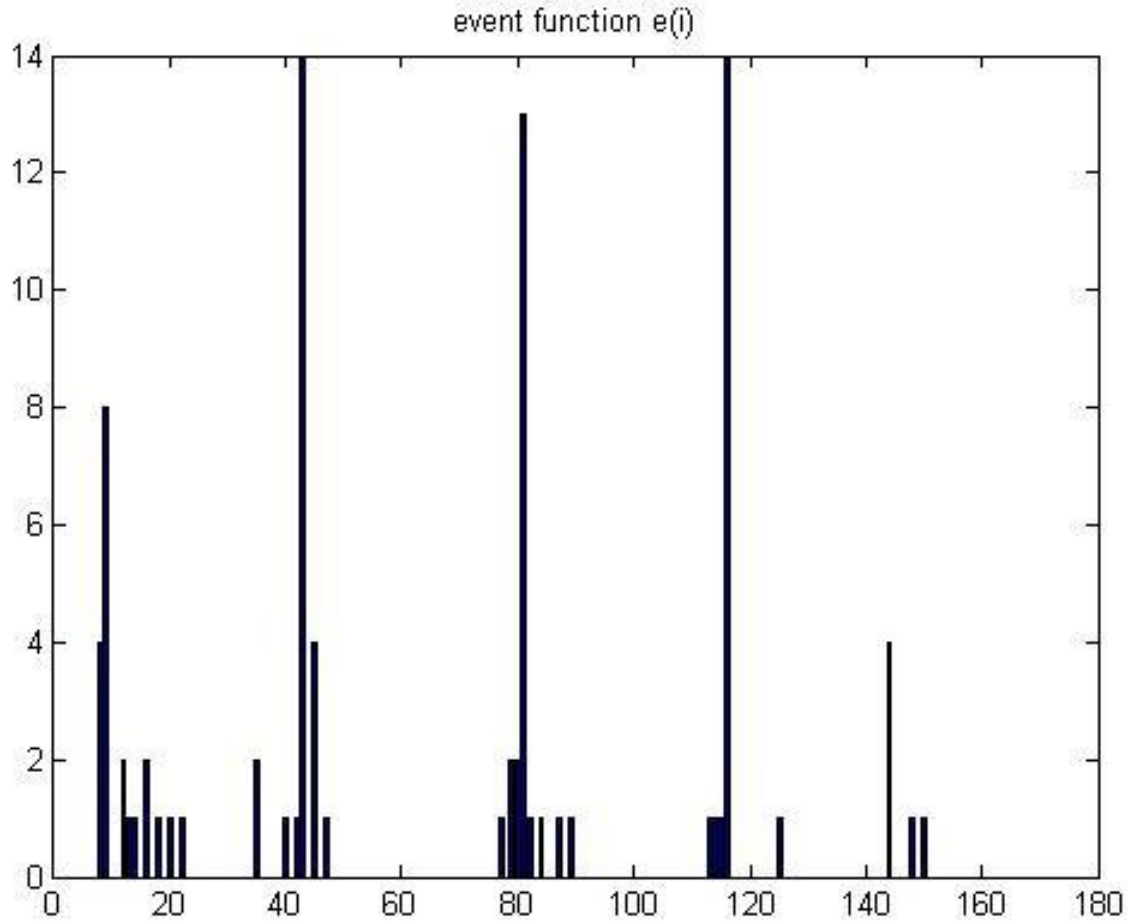
Events

Description	Mathematical condition	Importance			
Quasi-crossing point	$\lvert\beta\lvert r_n(i)\rvert - p'_n(i)\rvert < p \;\; AND$ $(\lvert\beta\lvert r_n(i+1)\rvert - p'_n(i+1)\rvert > p \;\; OR$ $\lvert\beta\lvert r_n(i-1)\rvert - p'_n(i-1)\rvert > p) \;\; AND$ $p'_n(i) > p_{min}$	1	3	4	1
Crossing points first case	$\lvert r_n(i)\rvert > p'_n(i) + p \;\; AND$ $\lvert r_n(i)\rvert < p'_n(i) - p \;\; AND$ $p'_n(i) > 5 * p_{min}$	1	3	4	1
Crossing points second case	$\lvert r_n(i)\rvert < p'_n(i) - p \;\; AND$ $\lvert r_n(i)\rvert > p'_n(i) + p \;\; AND$ $p'_n(i) > 5 * p_{min}$	1	3	4	1
Rate-of-change higher than power envelope	$\lvert r_n(i)\rvert > p'_n(i) \;\; AND$ $p'_n(i) > 2 * p_{min}$	1	2	2	1

Events


Quasi-crossing point
Derivative
p
Power
p
p min
n
n+1
Crossing
point 1. case
Power
p
p
5*p min
Derivative
n
n+1
Power
Crossing
point 2. case
p
Derivative
p
5*p min
n
n+1
Rate-of-change higher
than envelope
Derivative
Power
2*p min
n

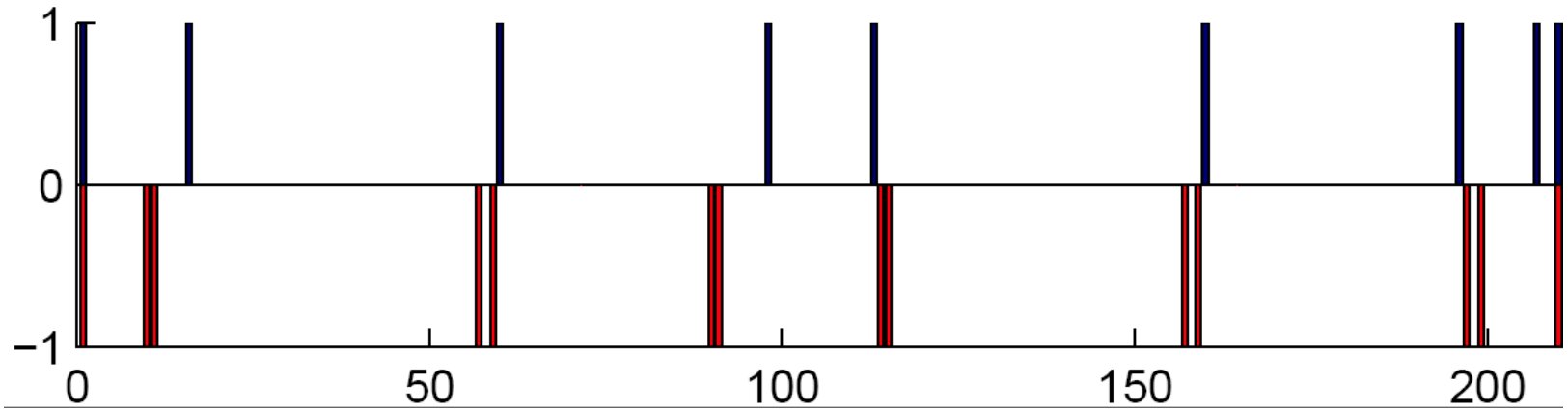
Event function



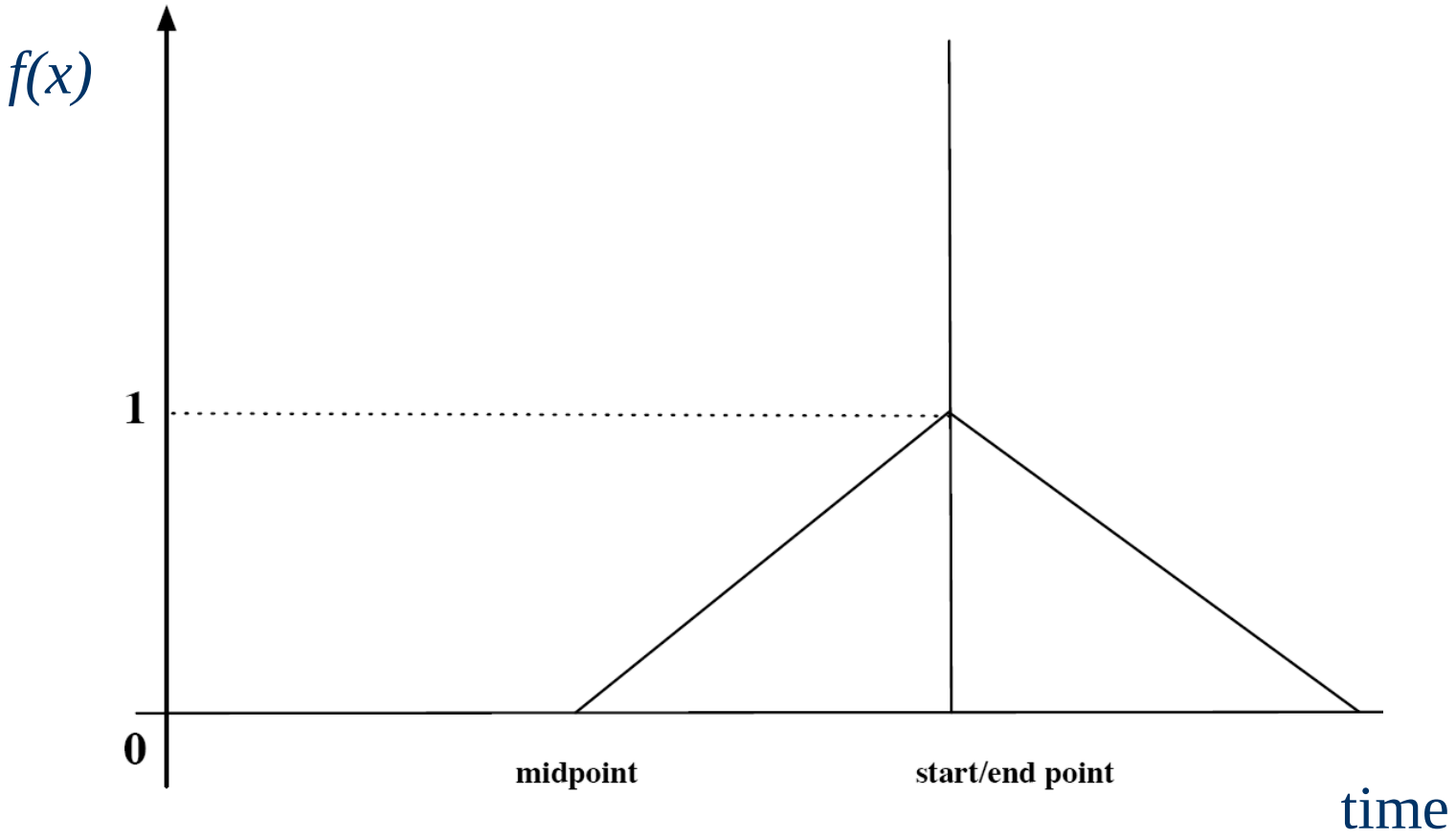
Approaches to speech segmentation evaluation

- Counting number of insertions, deletions and substitutions
- Counting the boundaries for which the deviation exceed thresholds (i.e. 35,70 ms)
- Value of tolerance is questionable
- Different inaccuracies treated in the same way
- Tolerances should be related to a length of an analysed phoneme

Example of segmentation



Fuzzy membership



Comparison

beg	9	56	89	113	156	196	-
end	10	58	90	114	158	198	-
auto	15	59	97	112	159	195	206
fuzzy recall and precision							
f(x)	0.78	0.93	0.36	0.91	0.95	0.95	0
insertations and deletions without tolerance							
Ins(7)	X	X	X	X	X	X	X
Del(6)	X	X	X	X	X	X	-
with tolerance from 1 (5.8 ms) to 4 (23.2 ms) - same results							
Ins(3)	X	✓	X	✓	✓	✓	X
Del(2)	X	✓	X	✓	✓	✓	-
with tolerance 5 (29 ms) or 6 (34.8 ms)							
Ins(2)	✓	✓	X	✓	✓	✓	X
Del(1)	✓	✓	X	✓	✓	✓	-
with tolerance 7 (40.6 ms) or higher							
Ins(1)	✓	✓	✓	✓	✓	✓	X
Del(0)	✓	✓	✓	✓	✓	✓	-

Results of segmentation

Method	av. recall	av. precision	F-score
Meyer	0.7096	0.7408	0.7249
db2	0.6770	0.7562	0.7144
db6	0.7029	0.7414	0.7217
db20	0.7034	0.7408	0.7216
sym6	0.7015	0.7426	0.7215
haar	0.6377	0.8042	0.7113
Meyer+sym6	0.6825	0.7936	0.7339
Meyer 7 subbands	0.6449	0.6714	0.6579

Method	av. recall	av. precision	F-score
Const 23.2 ms	0.9651	0.1431	0.2493
Const 92.8 ms	0.7635	0.4659	0.5787
SVM	0.50	0.33	0.40
Wavelet	0.7096	0.7408	0.7249

POS tagger

Part-of-speech (POS) tagging is the process of marking up the words as corresponding to a particular part of speech, based on both its definition, as well as its context, using their relationship with other words in a phrase, sentence, or paragraph. Many words represent more than one part of speech at different times.

The latest accuracy for Polish tagger is

Applying POS tagger in language modelling

331 occurrences were analysed. 282 of them had correct recognition in the whole 10 best list. Exactly 244 of all occurrences had a correct hypothesis on the 1 position of the 10 best list. 0.7372 % of occurrences were correctly recognised while using only HTK acoustic model. Only 53 occurrences were recognised applying probabilities from the POS tagger, even when HTK probabilities were 4 times more important than those from POS tagger. The weight was applied by raising HTK probability to power of 4. It gives 0.1601 % of correct recognitions.

$$p = p_{htk}^{w} p_{pos}.$$

Semantic model

Big John has a house. Big John has a black, aggressive cat. The black aggressive cat has a small mouse. The small mouse is a mammal.

All articles will be skipped as they have no semantic content and they do not exist in Polish which was our experimental language.

The first step of the algorithm is to count all other words in different topics, creating a matrix *S.*

Words-in-topics matrix S

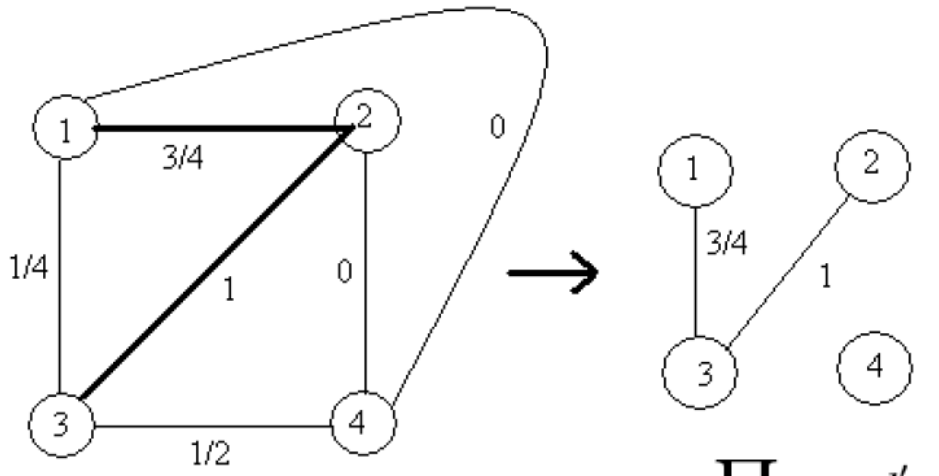
topic	big	John	has	house	black	aggr.	cat	small	mouse	is	mammal
1	1	1	1	1	0	0	0	0	0	0	0
2	1	1	1	0	1	1	1	0	0	0	0
3	0	0	1	0	1	1	1	1	1	0	0
4	0	0	0	0	0	0	0	1	1	1	1
3'	7/8	7/8	15/8	1/2	11/8	11/8	11/8	1	1	0	0

where rows $i = 1, \ldots, I$ represent topics and columns $k = 1, \ldots, K$ represent words. Each matrix value s_{ik} is the number of times word k occurs in topic i. A measure of similarity between two topics is

$$d_{ij} = \sum_{k=1}^{K} s_{ik} s_{jk}$$

$$d'_{ij} = d_{ij} / \max_{i<j} \{d_{ij}\}$$

Topic similarities



$$p_{ij} = \prod_{(a,b)\in P(i,j)} d'_{ab}$$

topic	1	2	3	4
1	4	3	1	0
2	3	6	4	0
3	1	4	6	2
4	0	0	2	4

where *P*(*i*, *j*) is the sequence of edges in the path from *i* to *j*. In the simplest case of a single edge *i* to *j* path d'_{ij} ight is . In case of multiple edges path, it is a product of similarities of all edges on a path.

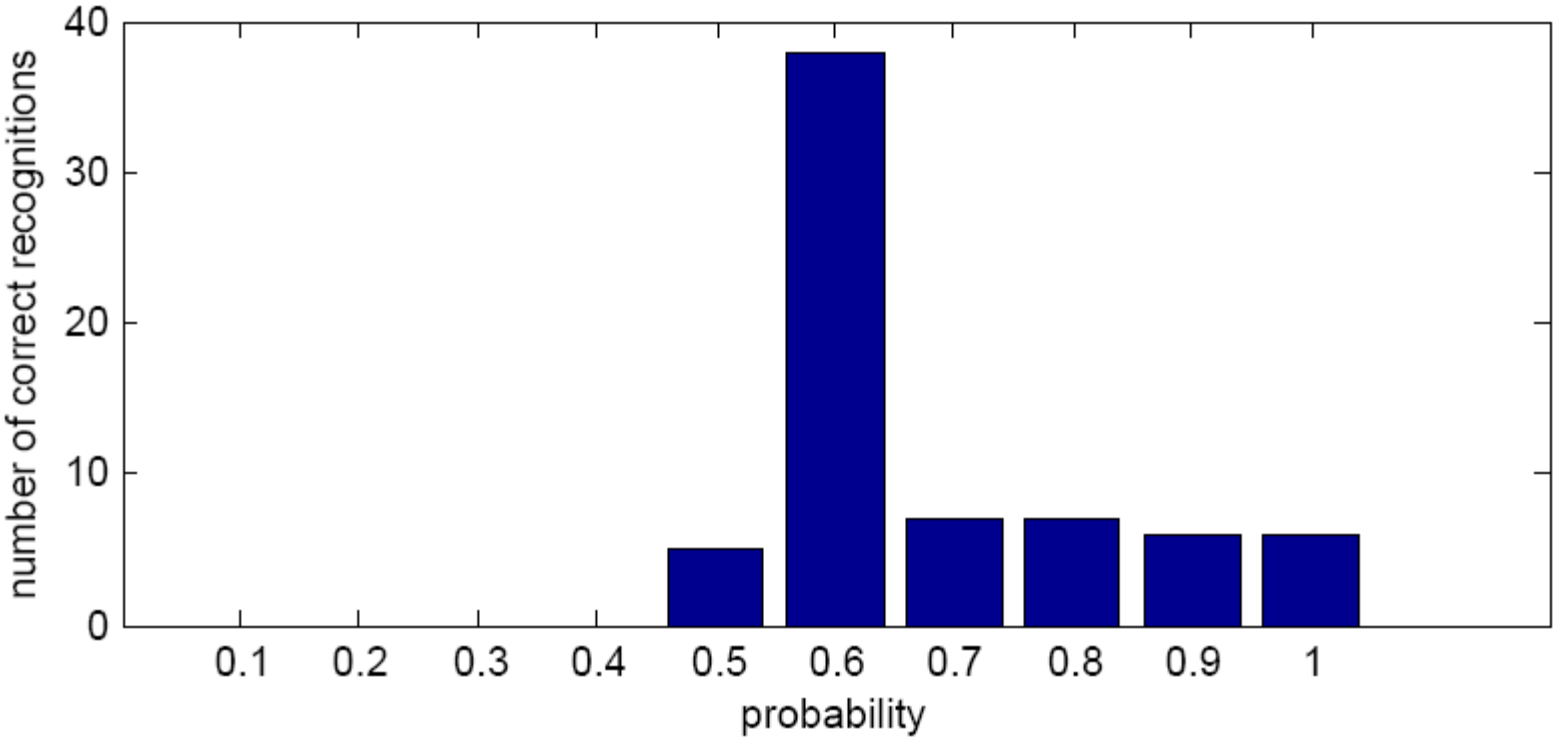
Modifying matrix S

For each node, we need to find *n* nodes with highest measures of paths leading to them from the given node. That will allow us to define a list *N* of semantically related topics which consists of the *n* nodes with their measures.

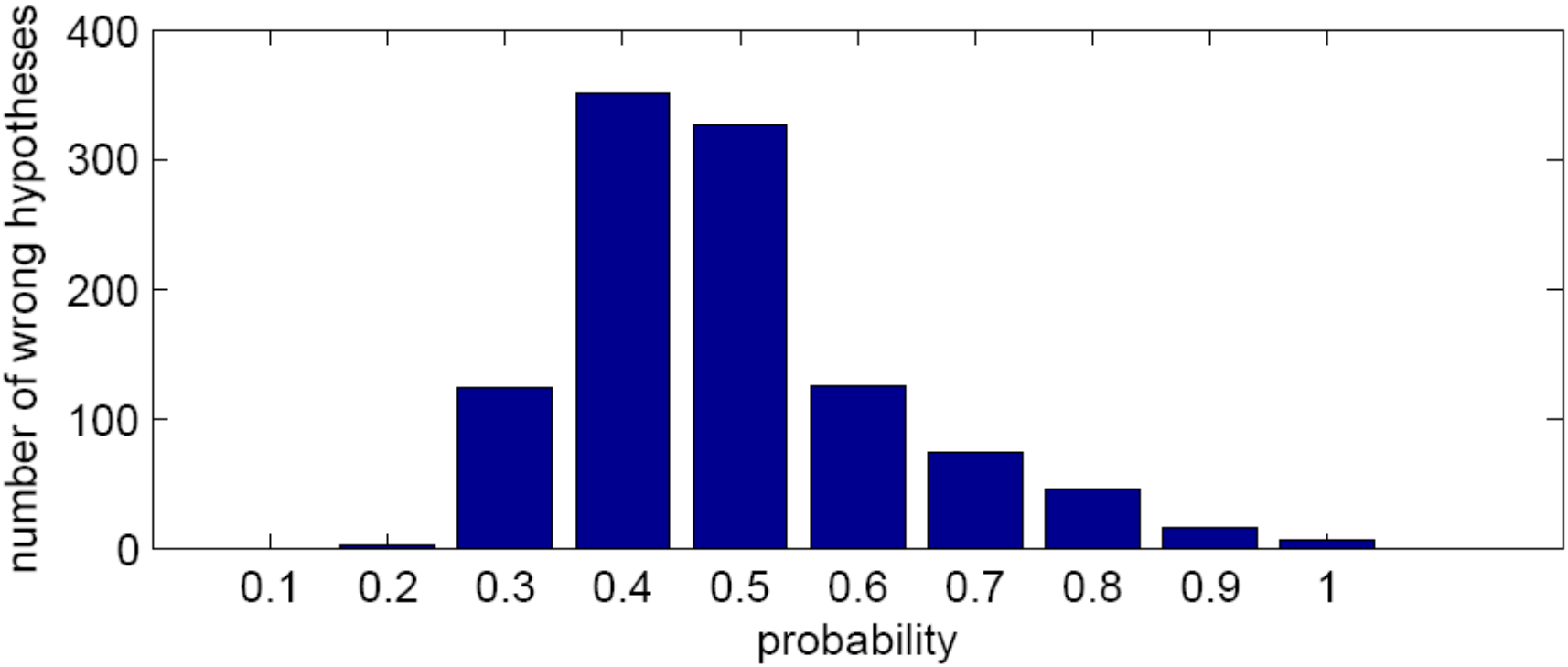
$$\mathbf{S}' = [s'_{ik}] \qquad s'_{ik} = s_{ik} + \alpha^{-1} \sum_{j \in N} p_{ij} s_{jk}$$

3’	7/8	7/8	15/8	1/2	11/8	11/8	11/8	1	1	0	0

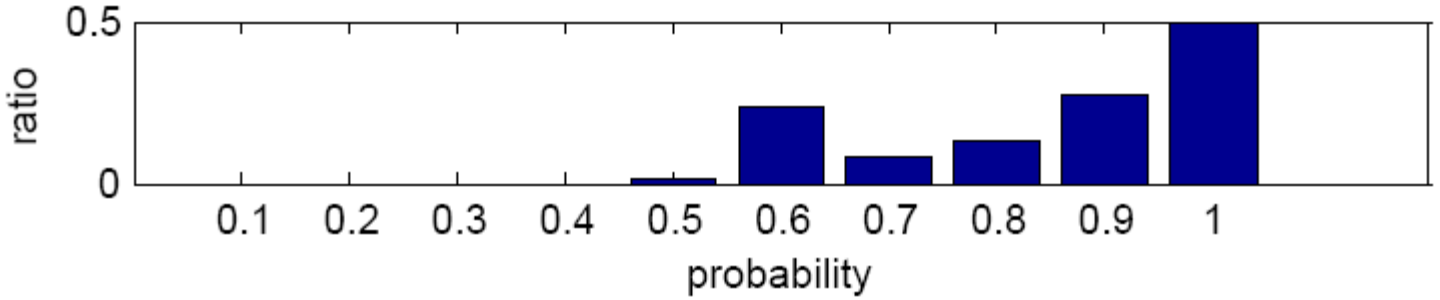
Histogram of correct recognitions



Histogram of wrong recognitions



Experimental results



HTK 29%

With LSA 36%

With our semantic model 53%

Recognition using semantic model

Recognition can be conducted by finding the most coherent topic for a set of words *W* in a provided hypothesis. It is carried on by finding a maximum of a sum of elements of *S'* from columns representing the word is over rows

$$P_{sem} = \max_i \sum_{k \in W} s'_{ik}$$

The row *i*, for which the maximum is found is assumed to represent the topic of sentence being recognised.

$$P_{sem} \in \Re^+ \qquad p = p^w_{htk} p_{sem}$$

Thank you!