Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 561–565

ISBN 978-83-60810-22-4
ISSN 1896-7094

Polish Phones Statistics

Bartosz Ziółko, Jakub Gałka,
Department of Electronics
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
www.dsp.agh.edu.pl
{bziolko,jgalka}@agh.edu.pl

***Abstract*—The phonemic statistics were collected from several large Polish corpora. The paper presents methodology of the acquisition process, summarisation of the data and some phenomena in the statistics. Triphone statistics apply context-dependent speech units which have an important role in speech technologies. The phonemic alphabet for Polish, SAMPA, and methods of providing phonemic transcriptions are described with detailed comments.**

***Index Terms*—Natural language processing, triphone statistics, speech processing, automatic speech recognition, Polish**

I. Introduction

The paper describes processing of linguistic data and constructing the statistics of Polish phone system by analysing large amount of text corpora using Cyfronet high performance computer cluster. There is a trade-off of quality of such statistics and time spent on calculations. The high performance computers enable obtaining the linguistic rules from the vast number of texts in reasonable time.

Statistical linguistics at the phone, word and sentence level are under considerations for several languages [2], [4], [9], [1]. The frequency of phonetic units appearance can be used in several speech processing applications, for example modelling in automatic speech recognition (ASR). Models of triphones which are not present in a training corpus of a speech recogniser can be prepared using phonetic decision trees [11]. The list of possible triphones has to be provided for a particular language along with phones' categorisation. The triphone statistics can also be used to generate hypotheses used in recognition of out-of-dictionary words including proper names or to provide additional probabilities in speech modelling (Fig. 1).

Some similar statistics collected from a few large corpora: Rzeczpospolita corpus (containing articles from a newspaper) [13], literature corpus [12] and Wikipedia corpus [14] (over 250 000 000 words) have been already presented. However, a space was included in the list of phones in the previous publications. This was not necessarilly a good decision because coarticulation happens between words as well. The process of speaking combines toegether several words and the borders between words are often indistinguishable on phonetic level.

Context-dependent modelling can significantly improve speech recognition quality. Each phone varies slightly depending on neighbouring phones due to a natural phenomena of coarticulation. There are no strict boundaries between phones because they overlap each other. It results in interference

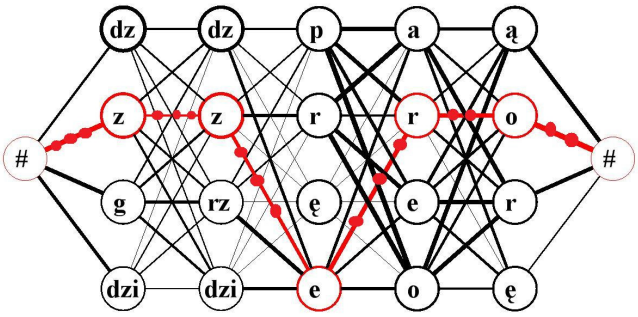

Fig. 1. An example of applying biphone statistics to speech recognition. The graph presents four phone hypotheses for each time slot with different probabilities (the highest row has the highest probabilities). The recognition based on audio information only would be *dz dz p a ąę*, which is not close to any Polish word. The best path after including biphone probabilities on edges is *z z e r o*, marked with extra dots. The word *zero* was actually spoken.

of acoustic properties. Speech recognisers based on triphone models rather than phone ones are much more complex but give better results [10]. Let us introduce examples of different ways of transcribing word *zero*. The phone model would be *z e r o* while the triphone one is **-z+e z-e+r e-r+o e-r+**. In case a specific triphone is not present, it can be replaced by a phonetically similar triphone (phones of the same phonetic group interfere in a similar way with their neighbours) using phonetic decision trees [11] or biphones (applying only left or right context) [8].

II. Algorithm

Sophisticated rules and methods are necessary to obtain the phonetic information from an orthographic text data. Transcription of text into phonetic data was applied by PolPhone [3] using extended SAMPA phonetic alphabet with 39 symbols (plus space) and pronunciation rules for cities Poznań and Kraków. Then the spaces were removed and our own digit symbols corresponding to SAMPA symbols were used, instead of typical ones in the aim of distinguishing phones easier while analysing received phonetic transcriptions. Stream editor (sed) was applied for this tasks.

Afterwards, statistics can be simply collected by counting the number of occurrences of each symbol, pair, and triple in an analysed phonetic transcription, where each character represents a phone. It was conducted using Matlab.

TABLE I
PHONES IN POLISH (SAMPA [3]), WHERE 1 % CORRESPONDS TO AROUND 11 190 000 OCCURENCES. THE LAST COLUMN PRESENTS THE RESULTS OBTAINED FROM MUCH SMALLER CORPUS A FEW DECADES AGO [6] (N.A. - NOT APLICABLE)

SAMPA	example	transcr.	%	[6]
a	pat	pat	9.584	9.3
e	test	test	9.108	10.2
o	pot	pot	8.994	9.1
t	test	test	4.489	4.4
r	ryk	rIk	4.674	3.6
n	nasz	naS	4.443	4
i	PIT	pit	4.359	3.9
j	jak	jak	3.796	4.5
I	typ	tIp	3.648	4.1
v	wilk	vilk	3.782	3.5
s	syk	sIk	3.638	3
u	puk	puk	3.345	3.4
p	pik	pik	3.263	3.1
m	mysz	mIS	2.988	3.5
k	kit	kit	2.976	2.7
d	dym	dIm	2.888	2.2
l	luk	luk	2.642	2.1
n'	koń	kon'	2.088	2.6
z	zbir	zbir	1.947	1.8
w	łyk	wIk	1.636	2.2
f	fan	fan	1.683	1.5
g	gen	gen	1.547	1.5
t^s	cyk	t^sIk	1.692	1.5
b	bit	bit	1.497	1.5
x	hymn	xImn	1.427	1.1
S	szyk	SIk	1.215	2
s'	świt	s'vit	0.965	1.5
Z	żyto	ZIto	0.944	1.2
t^S	czyn	t^SIn	0.955	1.2
t^s'	ćma	t^s'ma	0.662	1.3
w∼	ciąża	ts'ow∼Za	0.673	0.7
c	kiedy	cjedy	0.698	n.a.
d^z'	dźwig	d^z'vik	0.554	0.8
N	pęk	peNk	0.329	0.8
d^z	dzwoń	d^zvon'	0.261	0.2
J	giełda	Jjewda	0.260	n.a.
z'	źle	z'le	0.195	0.2
j∼	więź	vjej∼s'	0.112	0.1
d^Z	dżem	d^Zem	0.040	0

The necessity of investigating large text corpus pointed to the use of the Polish phonetic transcription system PolPhone [5], [3]. The transcription process is performed by a table-based system, which implements the rules of transcription. Matrix $T \in S^{m \times n}$ is a *transcription table*, where S is a set of strings and the cells meet the requirements listed precisely in [3]. The first element $t_{1,1}$ of each table contains currently processed character of the input string. For every character (or character substring) a table is defined. The first column of each table $\{t_{i,1}\}_{i=1}^{m}$ contains all possible character strings that could precede currently transcribed character. The first row $\{t_{1,j}\}_{j=1}^{n}$ contains all possible character strings that can proceed. All possible phonetic transcription results are stored in the remaining cells $\{t_{i,j}\}_{i=2,j=2}^{m,n}$. A particular element $t_{i,j}$ is applied as a transcription result, if $t_{i,1}$ matches the substring preceding $t_{1,1}$ and $t_{1,j}$ matches the substring proceeding $t_{1,1}$. The longer context is always prefered for transcription, to increase accuracy. Additional tables handle exceptions.

TABLE II
MOST COMMON POLISH BIPHONES. 1% CORRESPONDS TO AROUND 11 190 000 OCCURENCES. THE THIRD COLUMN PROVIDES INFORMATION ON AN INDEX OF A PARTICULAR BIPHONE IN ŁOBACZ AND JASSEM STATISTICS [7]

biphone	%	[7]	biphone	%	[7]
je	1.7253	1	ej	0.6620	13
ov	1.1829	12	do	0.6459	34
na	1.1632	3	or	0.6413	103
st	1.0791	7	ja	0.6367	5
po	1.0479	10	te	0.6229	9
ra	0.9189	14	ne	0.60803	57
ro	0.9155	21	em	0.60411	11
on	0.8756	18	at	0.60024	
n'e	0.8438	2	li	0.58227	68
ta	0.8035	4	to	0.58148	8
va	0.8012	33	re	0.5705	92
ar	0.7545	48	al	0.5654	35
ko	0.7337	25	aw	0.5595	32
er	0.7237	44	no	0.5410	19
an	0.6991	20	od	0.5386	71
en	0.6768	27	ka	0.54	39

TABLE III
THE REST OF MOST COMMON POLISH BIPHONES. 1% CORRESPONDS TO AROUND 11 190 000 OCCURENCES

biphone	%	biphone	%	biphone	%
eg	0.529	ek	0.445	vo	0.366
n'i	0.526	vj	0.442	ep	0.361
vy	0.526	in	0.434	ev	0.360
av	0.523	aj	0.427	et	0.356
go	0.515	pS	0.425	at^s	0.355
ow∼	0.506	ad	0.423	el	0.349
ty	0.506	tu	0.421	ym	0.345
za	0.497	op	0.419	Ze	0.345
ny	0.493	as	0.415	ve	0.342
os	0.489	ed	0.410	is	0.339
es	0.489	da	0.409	om	0.337
jo	0.488	t^se	0.401	wo	0.333
ol	0.485	mi	0.394	vi	0.332
am	0.480	ap	0.388	de	0.326
sp	0.473	ez	0.387	n'a	0.326
ma	0.470	nt	0.386	uv	0.321
pr	0.458	ku	0.383	az	0.321
Se	0.456	la	0.383	ok	0.316
en'	0.453	yx	0.378	s'e	0.310
le	0.449	ak	0.373	mo	0.307
an'	0.449	wa	0.370	ur	0.305
ci	0.447	ru	0.368	ob	0.304
ot	0.445	mj	0.368		

The phonetic alphabet used in PolPhone does not differentiate h and χ. However, it does differentiate w∼ and j∼. It can be seen as an unusual phonetical decision, but we are forced to use the existing tool as it is.

Several Rzeczpospolita (Polish daily journal) and Wikipedia articles were used as input data in our experiment. Due to their character, they contain quite many names and places, including foreign ones, what may influence the results slightly. The corpus consists also of several literature books in Polish. Some of them are translations from other languages, so they also contain foreign words. The whole corpus consists of around 267 000 000 words of over 3 000 000 word tokens.

TABLE IV
MOST COMMON POLISH TRIPHONES. 1% CORRESPONDS TO AROUND 11 190 000 OCCURENCES. THE THIRD COLUMN PROVIDES INFORMATION ON AN INDEX OF A PARTICULAR TRIPHONE IN ŁOBACZ AND JASSEM STATISTICS [7]

triphone	%	[7]	triphone	%	[7]
ova	0.3801	10	nyx	0.1673	
ego	0.3655	2	spo	0.1627	96
sta	0.3287	9	an'e	0.1586	16
vje	0.3159	1	pol	0.1538	
pSe	0.2969	6	os't^s'	0.1533	138
mje	0.2503	5	jej	0.1514	168
cje	0.2484		tur	0.1448	25
ovy	0.1942		jer	0.1433	86
jon	0.189	79	jow~	0.143	
ent	0.1842	76	ovj	0.1404	
pro	0.1807	41	ona	0.1381	38
ost	0.1785	19	ist	0.1371	204
ont^s	0.1749		en'e	0.1354	14
sci	0.1735		sto	0.1347	31
est	0.1734	8	an'a	0.1347	
ana	0.1722	21	ktu	0.1311	
ove	0.1712		ter	0.131	
pra	0.1681	33	s'c'i	0.130	

TABLE V
REST OF THE MOST COMMON POLISH TRIPHONES. 1% CORRESPONDS TO AROUND 11 190 000 OCCURENCES

triphone	%	triphone	%	triphone	%
jeg	0.130	ado	0.108	tra	0.0946
apo	0.130	ont	0.107	n'em	0.0941
nov	0.129	odo	0.106	era	0.0940
epo	0.129	any	0.106	n'ej	0.0938
jed	0.127	ora	0.106	jen'	0.0929
ajo	0.126	nt^se	0.106	end	0.0928
ast	0.124	ata	0.105	ano	0.0922
tov	0.124	ska	0.104	ejs	0.0922
van	0.123	pot	0.104	stf	0.0920
ina	0.122	neg	0.103	min	0.0912
pov	0.122	rat^s'	0.103	ami	0.0912
ali	0.122	awa	0.102	nte	0.0909
yst	0.121	oli	0.102	rek	0.0905
pje	0.120	tem	0.102	val	0.0902
ena	0.120	rav	0.1016	n'ik	0.0895
scj	0.120	rov	0.1009	avj	0.0892
pSy	0.118	nej	0.1008	gra	0.0890
dov	0.118	en'a	0.1002	ada	0.0886
ale	0.116	opo	0.0996	at^s'j	0.0885
ste	0.116	naj	0.0991	sko	0.0884
ovo	0.115	mja	0.0989	an'i	0.0881
van'	0.114	jen	0.0978	eta	0.0881
kon	0.113	ako	0.0976	jez	0.0876
tor	0.112	oku	0.0968	ate	0.0873
kov	0.110	art	0.0965	tan	0.0862
str	0.110	ane	0.0965	ama	0.0852
pod	0.108	rod	0.0964	oje	0.0849
zna	0.108	nym	0.0963	nap	0.0843

III. RESULTS

The frequency of phones is quite similar to the presented in [6] (Table I). The comparison to the other statistics [7] is not simple, because they include a space. It should be mentioned here, that it is an acoustic space, which we could rather call short pause. It means, that it does not appear between all words, but only where a speaker took a breath. A general correlation can be seen, however, the exact order of the statistics differs quite a lot (Tables II and IV).

The total number of around 1 119 000 000 phones were analysed with 39 tokens being specified. Exactly 1 397 biphonetokens (Fig. 2 and Table II) for 1 521 possible combinations were found, which constitutes 91.8%.

38 708 triphone tokens (see Table IV) were detected. The list of the most common triphones is presented in Table IV. With 39 phone tokens there are 59 319 possible triples. It leads to a conclusion that around 65% of possible triples were detected as triphones. It corresponds very well to Young [10], who estimates that in English, 60-70% of possible triples exist as triphones. It allows to make an assumption that all or nearly all triphone tokens were detected in our experiment.

Some values are similar to statistics given by Jassem a few decades ago and reprinted in [1]. We applied computer clusters, so our statistics were calculated for much more data. On the other hand, Jassem's work was based on manual transcription, while ours uses an automatic method to provide phone transcription, which might be less acurate.

Our results were compared with [7]. The phone statistics are quite similar, because they can be extracted from a small corpus and be representative. Because of this similarity we believe that the grapheme-to-phone automatic method we used has quality close to hand transcriptions done by [7]. The biphone statistics from [7] are less correlated to ours, with triphone ones quite different. We concluded that with correlation on uniphone level, the differences for biphones and triphones are due to much larger corpus which was analysed in our work. Biphone and triphone statistics from [7] are probably quite dependant on transcriptions which were used in their experiment. Our experiment was conducted on much larger corpus so it is much more data independent.

Besides the frequency of triphone occurrence, we are also interested in distributions of their frequencies. These are presented in logarithmic scale in Fig. 3. We have found around 2 200 triphones which occurred once, around 1 200 which occurred twice, and 900 three times. There are quite a lot of triples which occured few times but also there are generally much more triples than for a similar experiment with a space as a phone because there are new triples on words connections. Some threshold can be set and the rarliest triphones can be removed as errors caused by unusual Polish word combinations, acronyms, slang and other variations of dictionary words, onomatopoeic words, foreign words, errors in the process of introducing phone transcriptions and typographical errors in the text corpora.

Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. In the case of triphones it seems to be not the case. The changes in frequency between the common triphones are smaller then would be expected from Zipf's law, while the changes between rare triphones are larger. This type of distribution is very good from practical point of view. It can be estimated that triphones which are in the right part of the Fig. 3, where differences are very sharp are errors.

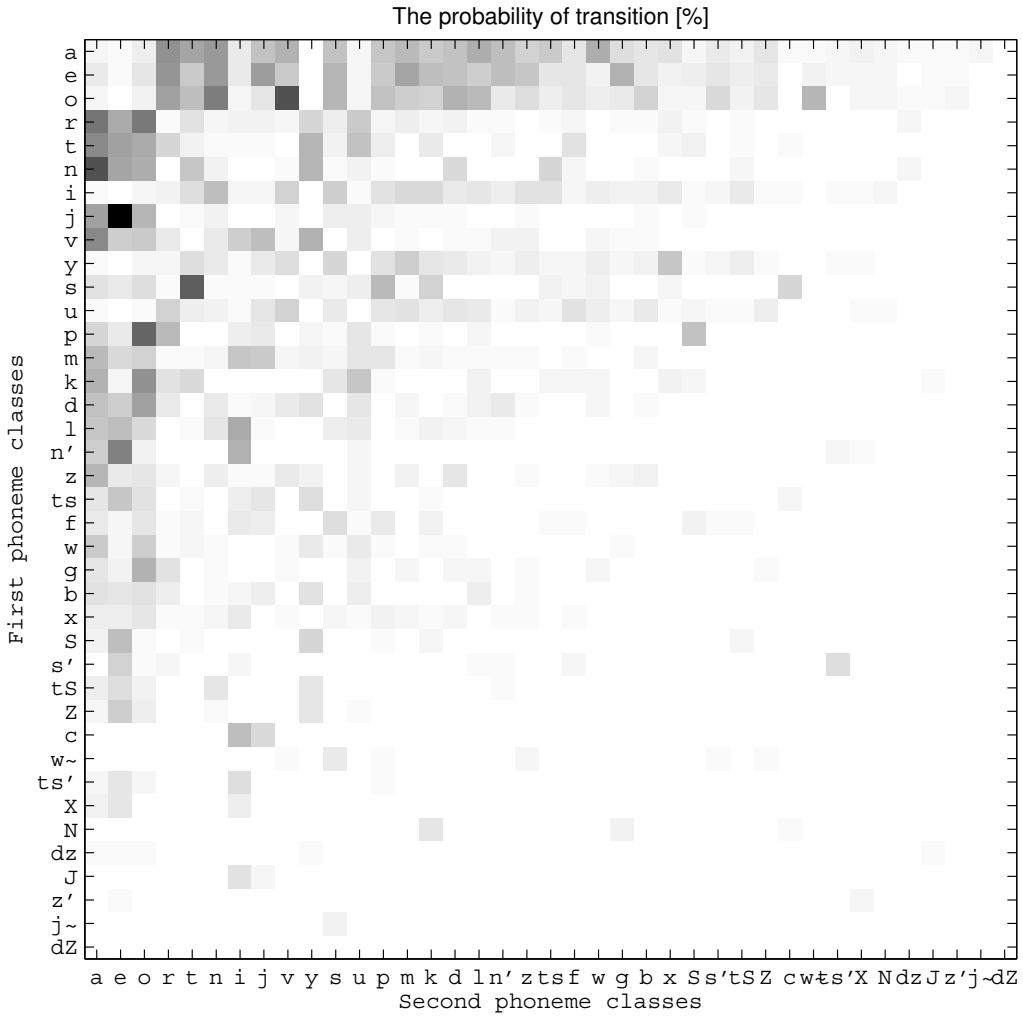

Fig. 2. Frequency of biphones in Polish

Spaces appear between written words but they rarely apear in spoken language. However, pauses can indeed occur between, e.g., longer phrases. Our statistics were collected using phonemic transciption based on written texts. One would say that a better way would be to use transcriptions made directly from audio records. However, in that case the amount of investigated material would be significantly smaller what would have impact on quality of the statistics, especially the triphones.

Entropy

$$H = -\sum_{i=1}^{40} p(i) \log_2 p(i), \tag{1}$$

where $p(i)$ is a probability of a particular phone, is used as a measure of the disorder of a linguistic system. It describes how many bits in average are needed to describe phones. According to Jassem in [1], the entropy for Polish is 4.7506 bits/phone but including a space. From our calculations, the entropy for phones is 4.7322, for biphones 8.6832 and 12.1987 for triphones. The fact that they do not follow a 1:2:3 proportion is an indication of some level of correlation.

IV. Conclusions

250 000 000 words from different corpora: newspaper articles, Internet and literature were analysed. Statistics of Polish phones, biphones and triphones were created. They are not fully complete, but the corpora were large enough, that they can be successfully applied in speech processing applications. The collected statistics are the largest for Polish, one of the most common Slavic languages, of this type of linguistic computational knowledge. It has several phones different than English and the statistics of phones are also different.

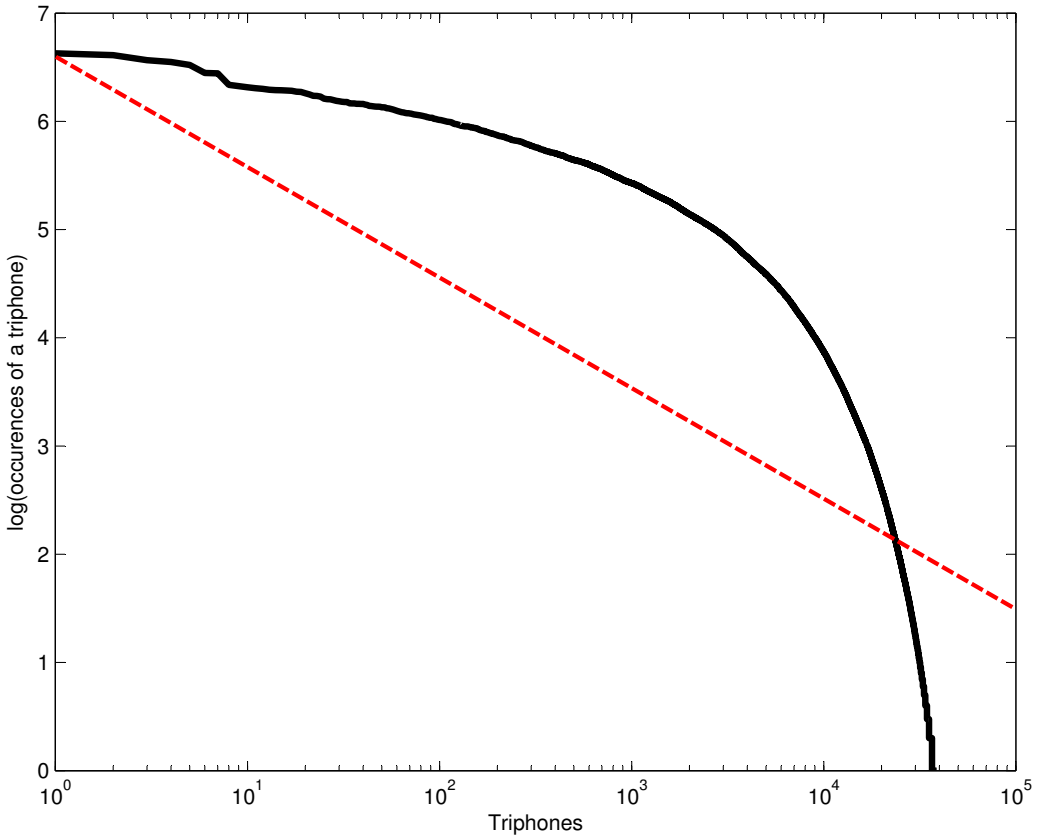

Fig. 3. The solid black line represents phone occurrences distribution while the red, dashed one is an ideal Zipf's law distribution (1/x). It can be estimated that the triphones which fall under the (1/x) Zipf's line are errors rather then real triphones and can be removed from the statistics

V. Acknowledgements

This work was supported by MNISW grant OR00001905. We would like to thank Institute of Linguistics, Adam Mickiewicz University in Poznań for providing PolPhone.

References

[1] C. Basztura, *Rozmawiać z komputerem (Eng. To speak with computers)*. Wrocław: Format, 1992.

[2] J. R. Bellegarda, "Large vocabulary speech recognition with multispan statistical language models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 76–84, 2000.

[3] G. Demenko, M. Wypych, and E. Baranowska, "Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis," *Speech and Language Technology, PTFon, Poznań*, vol. 7, no. 17, 2003.

[4] P. B. Denes, "Statistics of spoken English," *The Journal of the Acoustical Society of America*, vol. 34, pp. 1978–1979, 1962.

[5] K. Jassem, "A phonemic transcription and syllable division rule engine," *Onomastica-Copernicus Research Colloquium, Edinburgh*, 1996.

[6] W. Jassem, *Podstawy fonetyki akustycznej (Eng. Rudiments of acoustic phonetics)*. Warszawa: Państwowe Wydawnictwo Naukowe, 1973.

[7] P. Łobacz and W. Jassem, "Fonotaktyczna analiza mówionego tekstu polskiego," *B. Rocławski, Wybór materiałów do studiowania fonologii, fonetyki, fonotaktyki i fonostatystyki języka polskiego*, 1979.

[8] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. New Jersey: PTR Prentice-Hall, Inc., 1993.

[9] E. J. Yannakoudakis and P. J. Hutton, "An assessment of n-phoneme statistics in phoneme guessing algorithms which aim to incorporate phonotactic constraints," *Speech Communication*, vol. 11, pp. 581 – 602, 1992.

[10] S. Young, "Large vocabulary continuous speech recognition: a review," *IEEE Signal Processing Magazine*, vol. 13(5), pp. 45–57, 1996.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *HTK Book*. UK: Cambridge University Engineering Department, 2005.

[12] B. Ziółko, J. Gałka, and M. Ziółko, "Phone, diphone and triphone statistics for polish language," *13th International Conference on Speech and Computer SPECOM, St. Petersburg*, 2009.

[13] ——, "Phoneme ngrams based on a polish newspaper corpus," *WORLDCOMP, Las Vegas*, 2009.

[14] ——, "Phonetic statistics from an internet articles corpus of polish language," *International Joint Conference Intelligent Information Systems, Kraków*, 2009.