2010 Fifth International Multi-conference on Computing in the Global Information Technology

Speech Modelling Based on Phone Statistics

Bartosz Ziółko, Dawid Skurzok, Jakub Gałka, and Mariusz Ziółko
Department of Electronics
AGH University of Science and Technology
Kraków, Poland
{*bziolko, jgalka, ziolko*}*@agh.edu.pl*
www.dsp.agh.edu.pl

***Abstract*—The statistics of Polish phones, biphones and triphones were collected from several corpora. The paper presents summarisation of the data and some statistics phenomena including a distribution of frequency of biphones and triphones occurring. The model applying these statistics in speech recognition is presented as well.**

***Keywords*-phoneme statistics; triphone statistics; Polish; speech recognition; speech modelling;**

I. Introduction

There are two general types of methods in speech and language modelling: based on linguistic rules and statistical data. There are different opinions which of them is better. Typically it can be observed that mixed solutions provide the best results. In this paper some phone, diphone and triphone statistics for Polish, collected from vast amount of phonetically transcribed data are presented. Such statistics can be used in applications of speech processing, for example automatic speech recognition (ASR) systems. It is difficult to provide proper acoustic data for all possible triphones to represent as audio parameters. There are known methods to prepare models of triphones which did not appear in a training corpus of a speech recogniser. Data of other similar triphones and phonological similarities between different phones can be used [1]. It means, that the list of possible triphones has to be provided for a particular language. The uni/bi/triphone statistics can be also used to generate hypotheses in ASR system more accurately. They can be especially efficiently used in recognition of out-of-dictionary words like proper names.

We have already presented some similar statistics collected from various corpora: around 10 000 000 words of mainly spoken language (transcripts) [2], Wikipedia (97 000 000 words) [3], literature (68 000 000 words) [4] and everyday journal (104 000 000) [5]. Here we present data combined from these corpora and with an additional literature corpus of 949 000 000 words. It gives 1 248 000 000 words in total.

Obtaining of phonetic information from an orthographic text data is not straightforward. Transcription of text into phonetic data has to be applied first. In our work, PolPhone software [6] was used for this aim. It applies the extended

Table I
Polish phones in SAMPA standard [6], where 1% corresponds to around 11 190 000 occurences. The last column presents the results from [7].

SAMPA	our	example	transcription	%	[7]
#	#		#	17.10	4.60
e	e	test	test	8.12	10.61
a	a	pat	pat	7.91	9.55
o	o	pot	pot	7.52	7.98
j	j	jak	jak	3.46	4.37
n	n	nasz	naS	3.39	4.03
t	t	test	test	3.39	4.84
i	i	PIT	pit	3.39	3.39
I	y	typ	tIp	3.37	3.84
r	r	ryk	rIk	2.98	2.94
v	v	wilk	vilk	2.89	3.16
m	m	mysz	mIS	2.89	3.50
p	p	pik	pik	2.65	2.99
u	u	puk	puk	2.62	2.83
s	s	syk	sIk	2.54	2.79
d	d	dym	dIm	2.19	2.10
k	k	kit	kit	2.09	2.53
w	w	łyk	wIk	2.06	1.81
n’	3	koń	kon’	1.97	2.41
l	l	luk	luk	1.92	1.92
z	z	zbir	zbir	1.67	1.50
g	g	gen	gen	1.38	1.31
b	b	bit	bit	1.34	1.53
S	S	szyk	SIk	1.32	1.86
f	f	fan	fan	1.19	1.30
s’	5	świt	s’vit	1.16	1.64
Z	Z	żyto	ZIto	1.06	1.34
t^s	7	cyk	t^sIk	1.06	1.18
x	x	hymn	xImn	1.01	1.04
t^S	0	czyn	t^SIn	0.89	1.17
t^s’	8	ćma	t^s’ma	0.83	1.18
d^z’	X	dźwig	d^z’vik	0.68	0.70
w∼	2	ciąża	ts’ow∼Za	0.63	0.51
c	c	kiedy	cjedy	0.50	0.69
d^z	6	dzwoń	d^zvon’	0.24	0.22
z’	4	źle	z’le	0.22	0.22
N	N	pęk	peNk	0.21	0.1
J	J	giełda	Jjewda	0.14	0.08
j∼	1	więź	vjej∼s’	0.06	n.a.
d^Z	9	dżem	d^Zem	0.04	0.04

978-0-7695-4181-5/10 $26.00 © 2010 IEEE
DOI 10.1109/ICCGI.2010.25

SAMPA phonetic alphabet with 39 symbols (plus space). Pronunciation rules typical for cities Kraków and Poznań were chosen. For programming reasons we used our own single letter only alphabet corresponding to SAMPA symbols, instead of typical ones, to distinguish phonemes easier, while analysing received phonetic transcriptions. Stream editor (SED) was used to change original phoneme transcriptions into a set of letters and digits designed by us to simplify calculations. Statistics can be now simply gathered by counting number of occurrences of each phone, phones pair and triple in analysed text, where each phone is just a symbol (single letter or digit).

II. Text to Phonetic Transcription

Two main approaches are used for the automatic transcription of texts into phonemic forms. The classical approach is based on phonetic grammatical rules specified by a human [8] or machine learning process [9]. The second solution utilises graphemic-phonetic dictionaries. Both mentioned methods were used in PolPhone to cover typical and exceptional transcriptions. Polish phonetic transcription rules are relatively easy to formalise because of their regularity. The foreign words and abbreviations were treated by applying general grapheme-to-phoneme rules for Polish. Numerals were skipped.

The transcription process is performed by a table-based system, which implements the rules of transcription. Matrix $T[1..m][1..n]$ is a *transcription table* and its elements meet a set of requirements [6]. $T[1][1]$ of each table contains currently processed character of an input string. For every character (or a character substring) one table is defined. The first column of each table ($T[i][1]$, where $i = 1, ..., m$) contains all possible character strings that could precede currently transcribed character. The first row ($T[1][j]$, where $j = 1, ..., m$) contains all possible character strings that can follow a currently transcribed character. All possible phonetic transcription results are stored in the remaining cells of the tables ($T[2..n][2..m]$). A particular element $T[i][j]$ is chosen as a transcription result, if $T[i][1]$ matches the substring preceding $T[1][1]$ and $T[1][j]$ matches the substring proceeding $T[1][1]$. This basic scheme is extended to cover overlapping phonetic contexts. If more then one result is possible, then longer context is chosen for transcription, which increases its accuracy. Exceptions are handled by additional tables in the similar way.

Specific transcription rules were provided by a human expert in an iterative process of testing and updating rules. Text corpora used in the training process consisted of various sample texts (newspaper articles) and a few thousand words and phrases including special cases and exceptions.

III. Corpora and Statistics

Several Rzeczpospolita (Polish everyday journal) and Wikipedia articles were used as input data in our experiment.

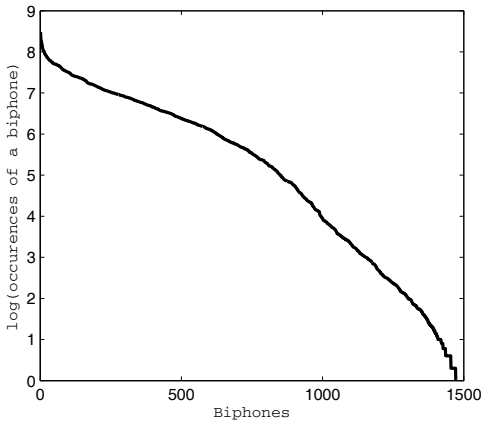

Figure 2. Distributions of biphones frequencies

Due to their character, they contain quite many proper names including foreign ones, what may influence the results slightly. The data consists also of several literature books in Polish. Some of them are translations from other languages, so they also contain foreign words. The whole collection consists of around 1 248 000 000 words.

Total number of 11 097 000 000 phonemes were analysed. They are grouped into 40 categories (including a space). Their distribution is presented in Table I. 1 473 different biphones (Fig.1 and Table II) for 1 600 possible combinations were found (92%). 38 182 different triphones (Table III) were found, where 64 000 are possible triples. It leads to a conclusion that around 60% of possible combinations actually exist as triphones, which is more than in our previous experiments [2], [3], [4], [5].

The most popular biphones (Table II) and triphones (Table III) were compared with the similar list provided in [7]. There are two important differences though. The first one is not very crucial for the list of most common combinations. Slightly different phonetic alphabets were used to cover some of the nuances of Polish. The second difference is much more crucial for the results. In our case, a space was between all words, as in the text files. In case of [7], the phonetic transcriptions were made based on voice recording, so the spaces were put only where a phonetician heard a space. It means that there are fewer spaces in [7] then in our statistics. As the result, combinations with spaces are rarlier.

The statistics for uniphones from our calculations are similar to [7]. The only main difference is in the frequency of a space. The reason for it was described in the paragraph above. The diphones statistics are quite similar. It can be counted that 19 of our 40 most popular biphones appeared between 40 most popular biphones in [7]. Similarly 79 of 118 biphones appeared in [7]. 104 of our 118 biphones appeared in the whole printed list in [7] (304 biphones). The

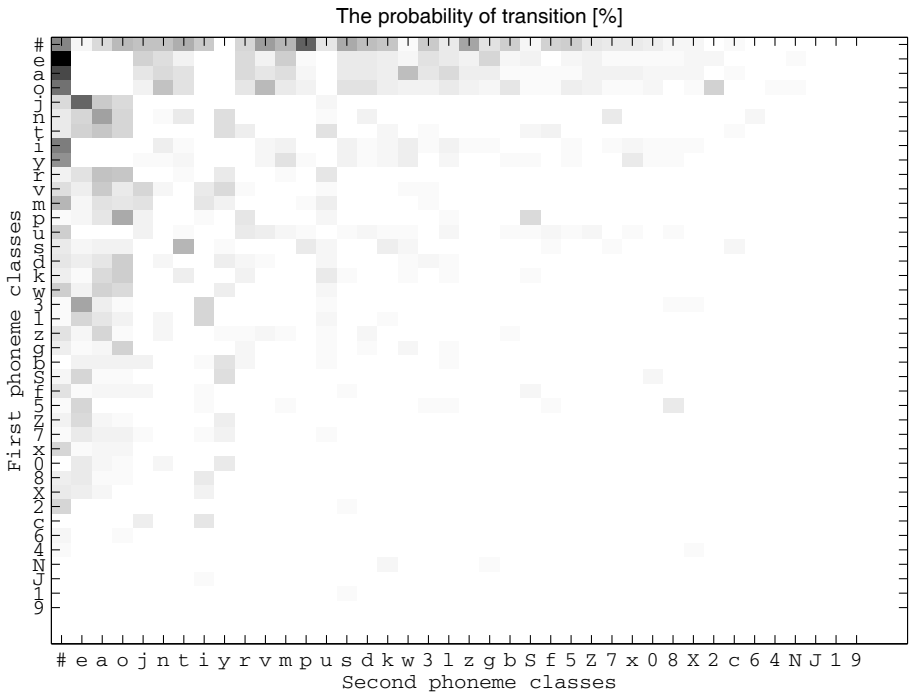

Figure 1. The distibution of diphones in Polish

statistics of triphones are far less similar, probably because the [7] experiment was conducted on quite little data so it is not representative. 11 of the 40 most popular triphones from [7] were found in our best 40. 37 of the best 116 from [7] appeared in the list of our 116 best. 57 of the best 304 (the whole printed list) of [7] were in the our list of the best 116. Both statistics have their flaws and advantages. Our statistics do not properly capture phonological nuances between words and has some bad elements due to errors inherited from Polphone. [7] has no such flaws but is based on very little data so the results for larger units are not representative enough.

Fig.1 shows some symmetry. It results from the fact that high values of α probability and β probability gives usually high probability of product $\alpha\beta$ and $\beta\alpha$ as well. Similar effects can be observed for triphones. Data presented in this paper illustrate the well-known fact that probabilities of triphones (Table III) cannot be calculated from the biphone probabilities (Table II). The reason for this is that the conditional probabilities have to be known.

Besides the frequency of particular combinations occurring, we are also interested in distributions of different frequencies, which are presented in logarithmic scale in Fig.3 (triphones) and Fig.2 (biphones). Some combinations of phones with very small occurrences are non-Polish and they should be excluded from the statistics. The rare biphones and triphones came from abbreviations, slang and

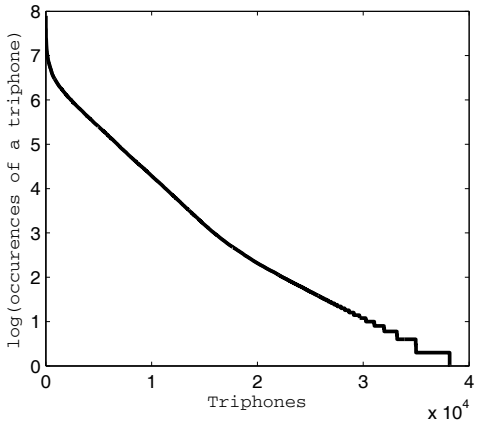

Figure 3. Distributions of triphones frequencies

non-dictionary words, onomatopoeic words, foreign words, errors in phonisation and typos in the text corpus.

Figs. 3 and 2 show also a well known phenomena in natural language processing that there is a small group of very common units and a large group of rare ones. Both functions start very sharply and then the rate of falling down decrease.

It can be assumed that from statistical point of view very rare units are not important, especially when smoothing

Table II
MOST POPULAR BIPHONES IN POLISH WITH THEIR POSITIONS IN A RANKING IN [7]

biphone	%	[7]	biphone	%	[7]
e#	2.755	6	j#	0.429	236
a#	1.961	22	ar	0.427	48
#p	1.703	n.a.	pS	0.425	46
je	1.674	1	Ze	0.420	15
o#	1.516	41	er	0.418	44
i#	1.401	91	al	0.409	35
y#	1.166	82	#a	0.409	90
#v	1.055	98	ka	0.409	39
na	0.999	3	wo	0.402	75
3e	0.983	2	vy	0.396	61
#z	0.973	172	an	0.392	20
#s	0.946	167	jo	0.388	136
po	0.921	10	en	0.377	27
#t	0.876	63	ny	0.366	n.a.
m#	0.809	102	am	0.365	17
#m	0.801	154	Sy	0.362	66
st	0.791	7	v#	0.355	n.a.
#o	0.756	153	ty	0.347	40
ov	0.736	12	#g	0.335	n.a.
aw	0.732	32	od	0.334	71
#d	0.708	206	ot	0.332	31
on	0.675	18	e3	0.330	43
#n	0.674	62	re	0.329	92
ra	0.664	14	f#	0.328	n.a.
#j	0.651	113	at	0.325	n.a.
ro	0.630	21	by	0.314	72
ta	0.626	4	ym	0.314	128
#k	0.607	217	mj	0.313	59
#i	0.599	143	os	0.311	291
ja	0.584	5	ma	0.308	58
va	0.570	33	z#	0.305	172
#3	0.567	110	tu	0.303	52
#b	0.555	195	#Z	0.299	261
ko	0.549	25	ru	0.296	107
w#	0.549	n.a.	aj	0.295	47
u#	0.548	139	#u	0.294	n.a.
#5	0.540	n.a.	ob	0.291	60
em	0.535	11	ci	0.290	245
do	0.523	34	mi	0.287	129
o2	0.510	56	d#	0.287	n.a.
go	0.495	29	or	0.285	103
te	0.493	9	av	0.284	54
#f	0.491	176	da	0.283	74
wa	0.491	42	pa	0.279	36
ej	0.478	13	a3	0.274	70
vj	0.468	16	mo	0.271	83
za	0.468	53	pr	0.270	87
to	0.464	8	la	0.263	130
5e	0.460	28	s#	0.254	n.a.
li	0.456	68	vi	0.253	94
x#	0.455	287	ed	0.250	67
no	0.453	19	#l	0.246	n.a.
ne	0.452	57	sp	0.244	96
Se	0.451	38	8i	0.242	122
eg	0.449	37	yx	0.240	126
le	0.448	24	el	0.239	141
2#	0.435	n.a.	ku	0.237	121
3i	0.433	65	ad	0.237	26
#r	0.433	n.a.	vo	0.236	149

Table III
MOST POPULAR TRIPHONES IN POLISH WITH THEIR POSITIONS IN A RANKING IN [7]. THE DIFFERENCES ARE MUCH MORE CRUCIAL THEN FOR PHONES AND BIPHONES

triphone	%	[7]	triphone	%	[7]
#po	0.712	37	#pa	0.161	n.a.
3e#	0.627	23	#vj	0.160	255
#na	0.571	26	on7	0.158	90
na#	0.459	56	awa	0.157	88
#3e	0.441	13	o#p	0.157	n.a.
o2#	0.435	n.a.	3a#	0.155	n.a.
#i#	0.402	n.a.	#z#	0.155	n.a.
#5e	0.401	n.a.	pSy	0.155	17
#za	0.399	291	e#m	0.154	n.a.
5e#	0.388	n.a.	#sp	0.149	n.a.
#pS	0.385	116	i#p	0.145	n.a.
go#	0.380	67	wo#	0.143	n.a.
#do	0.371	195	#ty	0.143	n.a.
ej#	0.359	132	ovj	0.141	15
je#	0.356	27	ale	0.137	4
vje	0.356	1	ka#	0.134	233
#je	0.353	123	pov	0.132	n.a.
ego	0.340	2	ost	0.128	19
em#	0.339	63	e#o	0.127	n.a.
aw#	0.279	n.a.	tur	0.126	25
e#p	0.274	210	8i#	0.126	n.a.
wa#	0.272	n.a.	e#t	0.126	102
#vy	0.265	228	ktu	0.125	30
pSe	0.261	6	jed	0.125	78
ova	0.255	10	ajo	0.124	n.a.
sta	0.250	9	o58	0.124	138
Ze#	0.245	92	pra	0.123	33
#v#	0.221	n.a.	#kt	0.122	n.a.
yx#	0.214	n.a.	pje	0.121	49
ym#	0.213	n.a.	e#d	0.120	n.a.
#f#	0.212	n.a.	jo2	0.119	n.a.
cje	0.211	n.a.	wy#	0.119	n.a.
#st	0.209	n.a.	#mj	0.119	n.a.
#Ze	0.205	115	ko#	0.118	n.a.
#ja	0.203	46	by#	0.118	n.a.
mje	0.202	5	vaw	0.117	277
ne#	0.196	164	#s#	0.116	n.a.
a#p	0.195	n.a.	a#v	0.116	n.a.
e#z	0.193	n.a.	jej	0.116	168
do#	0.191	n.a.	dy#	0.115	n.a.
#te	0.191	220	#mu	0.115	n.a.
to#	0.187	201	#a#	0.115	n.a.
#to	0.187	34	byw	0.114	51
e#v	0.187	n.a.	8e#	0.114	n.a.
jon	0.186	79	jeg	0.114	296
#ko	0.182	n.a.	bje	0.114	75
ny#	0.180	179	#3i	0.113	n.a.
li#	0.175	n.a.	e#n	0.112	124
#ro	0.173	n.a.	y#p	0.111	n.a.
#pr	0.171	n.a.	58i	0.109	66
#by	0.170	n.a.	jer	0.107	86
mi#	0.167	n.a.	ent	0.107	76
le#	0.163	297	e#j	0.106	249
ci#	0.163	271	a#t	0.105	n.a.
e#s	0.162	n.a.	3i#	0.105	n.a.
#ta	0.161	n.a.	#od	0.104	n.a.
#mo	0.161	n.a.	a#s	0.103	n.a.
#ma	0.161	n.a.	e#b	0.103	n.a.

operation is applied in order to eliminate disturbances caused by lack of text data. All units which appeared less then 10 times can be considered as errors. This is why we should recalculate values from the above paragraph.

There were 1 408 biphones which appeared more then 10 times. With 1 600 possible combinations it gives 88% of possible combinations found as biphones. There were 30 251 triphones which appeared more then 10 times. There are 64 000 possible triples. It results in 47% of possible combinations which actually exist as 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 [10], the entropy is 4.7506 bits/phone. From our calculations entropy for phones is 4.5993, for biphones 8.2995 and 11.4763 for triphones.

IV. Model Description

N-grams model is probably the model which is the most often applied for language modelling. However, the same mathematical tool can be applied for sequences of phones.

Context-dependent units can improve recognition highly. In speech, applying biphones and triphones gives this opportunity. Phonemes vary slightly depending on the neighbouring ones due to a natural phenomena of coarticulation. There are no clear boundaries between phonemes. They rather overlap each other, which results in starts and ends being dependant on other phonemes. Speech recognisers based on triphone models rather than phoneme ones are much more complex but give better results [1]. Examples for transcribing *above* are: *ax b ah v* (a phoneme model) and **-ax+b ax-b+ah b-ah+v ah-v+** (a triphone model). In case a specific triphone is not present there are two options. It can be synthesised using phonetically similar triphones because phonemes of the same phonetic group interfere in similar way with their neighbours. Another approach is to estimate it with biphones, applying left and right context separately.

Our model searches for the shortest path through a graph like the one presented in Fig. 4 using Dijkstra algorithm [11] with a small modification, which is a result of the fact that a triphone statistical probability corresponds to two arches at once, not one. Four different probabilities are included in the calculations. First two are related to nodes: the probability of an acoustic hypothesis from a classifier and probability of a phone from statistics. Two others are weights on arches: statistical probability of biphones and triphones. The triphone statistics are used in a way that only a locally best path for each node to every following node is kept and considered. It is calculated using all possible combinations of triphones for the particular node.

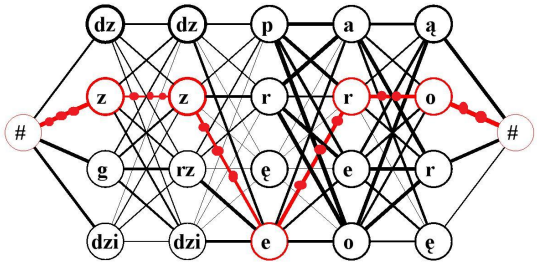

Figure 4. An example of applying the biphone statistics to speech recognition. The graph presents four phone hypotheses for each time slot with different probabilities (the highest row has the highest probabilities). The recognition based on audio information only would be *dz dz p a ą*, which is not 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.

Table IV
Comparison of recognition using only a classifier and with applying a model based on phonetic statistics. Results are presented as an average distance between the best hypothesis and the spoken word; the average position of the spoken word in the ranking of all classified words from the dictionary (around 1 200) and as a percentage of spoken words for which the proper recognition appeared in the list of the 5-best hypotheses.

method	ed. metric	av. position	% in 5-best
ground truth	12.5	182.4	34.2
statistical model	11.8	156.8	39.2

V. Results

Our ASR classifier was trained on 20 of the male speakers from CORPORA. Another male speaker, which was not included in the training, was used for tests. However the same words were recorded. The model was tested for 1216 words. 1000 words were added to the dictionary (apart from words from CORPORA) to make recognition task more difficult. Results are presented in Table IV in three ways: average distance in an edit metrics, average position of the correct word in the list of the hypotheses, percentage of words which appeared in the 5-best list of the hypotheses. A medium dictionary was used with no language modelling. This is why a word error rate would be too low to observe behaviour of the model comparing to the ground truth.

VI. Conclusions

Phone, biphone and triphone statistics for Polish were compared with existing literature. A model based on the statistics improved results for around 10% for words from a dictionary. It can be especially useful for recognition of out-of-vocabulary words, were it can form a hypothesis in a readable and human friendly version based on phonetic statistics.

Presented statistical model was implemented in our medium vocabulary, continuous speech recognition software, which is a prototype of a developing Polish large vocabulary, continuous speech recognition system.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] 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.

[2] B. Ziółko, J. Gałka, S. Manandhar, R. Wilson, and M. Ziółko, "Triphone statistics for Polish language," *Proceedings of 3rd Language and Technology Conference, Poznań*, 2007.

[3] B. Ziółko, J. Gałka, and M. Ziółko, "Phonetic statistics from an internet articles corpus of polish language," *International Joint Conference Intelligent Information Systems, Kraków*, 2009.

[4] ——, "Phone, diphone and triphone statistics for polish language," *13th International Conference on Speech and Computer SPECOM, St. Petersburg*, 2009.

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

[6] 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.

[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] M. Steffen-Batóg and P. Nowakowski, "An algorithm for phonetic transcription of ortographic texts in Polish," *Studia Phonetica Posnaniensia*, vol. 3, 1993.

[9] W. Daelemans and A. van den Bosch, "Language-independent data-oriented grapheme-to-phoneme conversion," *Progress in Speech Synthesis, New York: Springer-Verlag*, 1997.

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

[11] E. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.