Phone, diphone and triphone statistics for Polish language

Bartosz Ziółko, Jakub Gałka, Mariusz Ziółko

Department of Electronics, AGH University of Science and Technology
al.Mickiewicza 30, 30-059 Kraków, Poland
{bziolko,jgalka,ziolko}@agh.edu.pl

Abstract

The statistics of Polish phonemes, diphones and triphones were collected from a large literature corpus. The paper presents summarisation of the data and focuses on interesting phenomena in the statistics. Triphone statistics play an important role in speech recognition systems. They are used to improve the proper transcription of the analysed speech segments. A distribution of frequency of triphones occurring and other phenomena are discussed. SAMPA - the standard phonetic alphabet for Polish and methods of providing phonetic transcriptions are described.

1. Introduction

Statistical research at the word and sentence level are popular for several languages [1, 2]. Any similar research on phonemes is rare [3, 4, 5]. The frequency of phonetic unit presence is an interesting topic itself. It can also be used in many applications in speech processing, for example speech recognition. It is very difficult to provide proper acoustic data for all possible triphones to represent them with audio parameters. There are methods to synthesise triphones which not appeared in a training corpus of a speech recogniser, using data of other triphones and phonological similarities between different phonemes [6]. It means, that the list of possible triphones has to be provided for a given language. The triphone statistics can be also used to generate hypotheses used in recognition of out-of-dictionary words like names.

We have already presented some similar statistics [7], which were collected from around 10 M words of mainly spoken language. Here we present statistical data collected from 61 M words from a corpus containing various literature books (originally Polish or translations into Polish). It will let us compare these statistics to evaluate how representative and complete they are. We have found in research on semantics [8] that often it is better to use formal written language corpus, rather then speech transcriptions to train speech models. Even though, it is counter-intuitive, the language from literature is always of better quality so it is a better source of linguistical knowledge. Speech transcriptions can be too random to help in representing any linguistical phenomena.

This paper describes several issues related to phoneme, diphone and triphone statistics and is divided as follows. Section 2 provides information about general scheme of our data acquisition method and standards we used. Section 3 describes the technically most difficult step which is changing the text corpus into a phonetic transcription. Section 4 contains a description of data we used and our results. Phenomena we uncovered are described as well. We sum up the paper with conclusions.

Table 1: Phoneme transcription in Polish - SAMPA [9]

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
J	giełda	Jjewda	507,679	0.12
dZ	dżem	dZem	201,248	0.05
j∼	więź	vjej∼s’	154,452	0.04

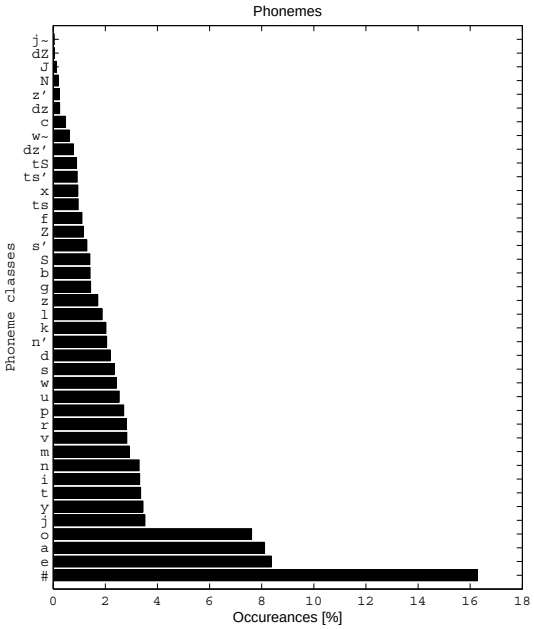

Figure 1: Phonemes in Polish

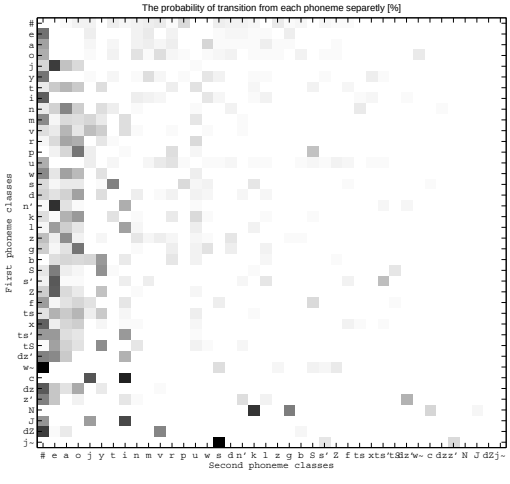

Figure 2: Diphone probabilities in Polish for each phoneme separately

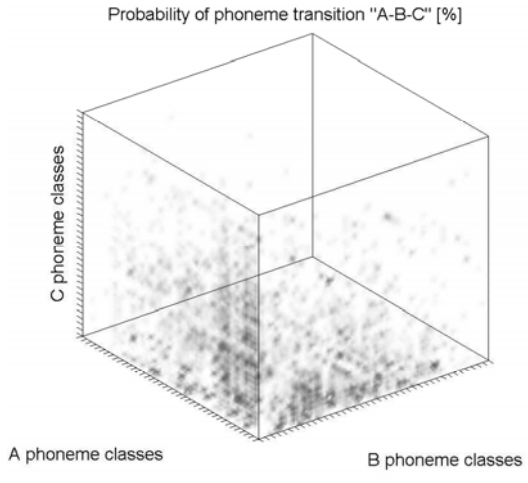

Figure 3: Triphone probabilities in Polish

2. General Description of the method

Obtaining of phonetic information from an orthographic text-data is not straightforward [10, 11]. Transcription of text into phonetic data has to be applied first [12]. We used PolPhone [9] software for this aim, which is described in the next section. The SAMPA extended phonetic alphabet was applied with 39 symbols and pronunciation rules typical for cities Kraków and Poznań. For programming reasons we used our own single letter only symbols corresponding to SAMPA symbols, instead of typical ones, to distinguish phonemes easier while analysing received phonetic transcriptions. SED was used to change original phoneme transcriptions into ours. Statistics can be now simply calculated by counting number of occurrences of each phoneme, phoneme pair, and phoneme triple in analysed text, where each phoneme is just a symbol. The last, analysing, part was conducted in Matlab on a high performance computer in the Academic Computer Centre CYFRONET AGH. Mars, the computer used, has following specification: IBM Blade Center HS21 - 112 Intel Dual-core processors, 8GB RAM/core, 5 TB disk storage and 1192 Gflops. It operates using Red Hat Linux.

3. Text to Phonetic Transcriptions

Two main approaches are used for the automatic transcription of texts into phonemic form. The classical approach is based on phonetic grammatical rules specified by human [13] or automatic machine learning process [14]. A second solution utilises graphemic-phonetic dictionaries. In practice, both mentioned methods are used in order to cover typical and exceptional transcriptions. Polish phonetic transcription rules are relatively easy to formalise because of their regularity.

The necessity of investigating large text corpus pointed to the use of the Polish phonetic transcription system PolPhone [15, 9]. In this system, strings of Polish characters are converted into their phonetic SAMPA representation. Extended SAMPA (Table 1) is used, to deal with all nuances of Polish phonetic system. The transcription process is performed by a table-based system, which implements the rules of transcription. A matrix $T[1..m][1..n]$ is a *transcription table* and its cells meet a set of requirements [9]. The first element ($T[1][1]$) of each table contains currently processed character of the input string. For every

character (or 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 preceed 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 (SAMPA symbols) 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 proceeding $T[1][1]$ and $T[1][j]$ matches the substring following $T[1][1]$. 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 manner.

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

4. Corpus and Results

Several literature books in Polish were used as input data in our experiment. Some of them are translations from other languages, so they can contain foreign names, what may influence the results slightly. In total, 490 Mbytes (around 61 MWords) were included in the proccess.

Total number of 417,128,060 phonemes were analysed. They are grouped in 40 categories (including space). Their distribution is presented in Table 1 and in Fig. 1. 1,151 different diphones (Fig. 2 and Table 2) and 16,864 different triphones (Fig. 3) were found. It has to be mentioned that all combinations like *#*, where * is any phoneme and # is space, were removed as we do not treat these triples as triphones. The reason for it, is that first phoneme * and the second one are actually in 2 different words, while we are interested in triphone statistics inside words. The list of 120 most common triphones is presented in Tables 3 and 4. Assuming 40 different phonemes (including space) and subtracting mentioned *#* combinations, there are 62,479 possible triples. We found 16,864 different triphones, what gives a conclusion that around 27% of possible combinations were actually found as triphones, which is a similar result to our previous experiment [7]. An average length of words in phonemes can be estimated as around 6, due to space (noted as #) frequency 16.28.

Besides the frequency of triphones occurring, we are also interested in distributions of different frequencies, which is presented in logarithmic scale in Fig. 4. We expected to receive another distribution, for which it would be easy to set a threshold between proper triphones and errors, as very large amount of text was analysed. We hoped to have very few triphones which occurred fewer than 3 times. Then we would deduce that they are not real triphones but errors due to foreign names etc. in the corpus. We noted around 500 triphones which occurred just once, 300 with occurrence 2, 200 with 3 to 6 occurrences, 100 for 7 to 9, and 50 for 10 or more. Such phenomena is nothing unexpected in natural language processing on a level of words or above, where amount of analysed text do not change statistics (considering reasonable large amounts). Still, in case of triphones, the number of possibilities is much smaller and limited to mentioned 62,479. We observed the same type of scenario in our previous experiment [7], however, this time we have fewer rare triphones with larger corpus. It supports a hy-

Table 2: Most common diphones in the analysed corpus

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
vj	2,117,396	0.508
za	2,114,797	0.507
te	2,044,260	0.490
le	2,014,379	0.483
Ze	2,002,119	0.480
to	1,992,362	0.478
Se	1,954,567	0.469
li	1,911,724	0.458
ej	1,910,533	0.458
no	1,902,527	0.456
#f	1,901,677	0.456
wo	1,879,265	0.450
n'i	1,871,643	0.449
eg	1,846,163	0.443
w∼#	1,843,488	0.442
pS	1,829,138	0.439

Table 3: Most common triphones in the analysed corpus

triphone	no. of occurrences	percentage
#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
mi#	735,868	0.176
#vj	723,174	0.173
ny#	718,709	0.172
#ma	712,084	0.171
wo#	711,870	0.171

Table 4: Most common triphones (2nd part)

triphone	no. of occurrences	percentage
#ko	684,104	0.164
e#t	676,933	0.162
#sp	673,546	0.161
yx#	671,756	0.161
#ro	670,581	0.161
ovj	670,429	0.161
ale	659,997	0.158
i#p	659,100	0.158
pov	655,294	0.157
#z#	644,656	0.157
onts	637,551	0.153
#ty	626,755	0.150
vaw	618,539	0.148
#pr	614,475	0.147
#mu	604,432	0.145
by#	598,726	0.144
e#o	598,241	0.143
am#	597,426	0.143
#n'i	584,630	0.140
ci#	575,219	0.138
byw	574,832	0.138
jed	570,810	0.137
e#d	568,404	0.136
e#j	567,051	0.136
e#n	552,936	0.133
iw#	547,333	0.131
ost	545,770	0.131
wy#	538,702	0.129
ts'e#	535,134	0.128
dy#	534,589	0.128
#tso	534,567	0.128
y#p	533,886	0.128
pje	533,518	0.128
bje	531,366	0.127
ko#	530,153	0.127
ka#	530,063	0.127
a#t	529,121	0.127
e#n'	523,839	0.126
ajo	517,881	0.124
#mj	512,467	0.123
#kt	512,135	0.123
e#b	508,620	0.122
pra	500,445	0.120
o#t	491,424	0.118
n'i#	483,610	0.116
#a#	482,164	0.116
jej	480,268	0.115
n'a#	479,154	0.115
a#v	478,479	0.115
#al	478,210	0.115
#s#	475,722	0.114
ktu	475,605	0.114
#ot	473,976	0.114
jeg	473,192	0.113
tur	470,613	0.113
my#	470,452	0.113
wem	468,184	0.112
a#z	464,227	0.111
a#m	463,180	0.111

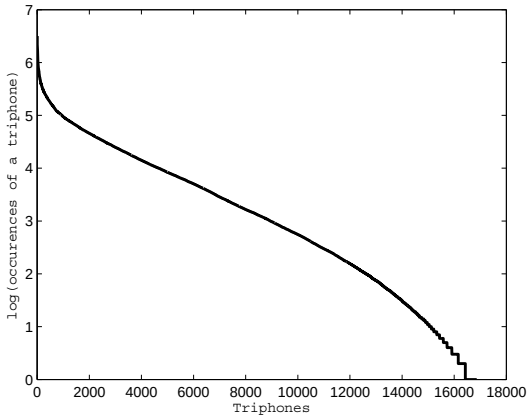

Figure 4: *Distribution of frequency of occurring phonemes in logarithmic scale*

pothesis that one can reach a situation, when new triphones do not appear and distribution of occurrences is changing as a result of more data being analysed. It is possible that the large number of triphones with very small occurrence are non-Polish triphones which should be excluded. The rare triphones come from unusual Polish word combinations, slang and other variations of dictionary words, onomatopoeic words, foreign words, errors in phonisation and typos in the text corpus. In our further works we will assume that from statistical point of view it is not important, especially when smoothing operation is applied in order to eliminate disturbances caused by lack of text data [16].

We observed that all statistics, even the phoneme one (Table 1 and Fig. 1), are quite different in this experiment then in the previous one [7]. We used a slightly different version of SAMPA alphabet there, but the differences between experiments, in order of phonemes can be easily spotted. In [7] phonemes were ordered by frequency in the list: a, e, o, s, t, r, p, v, j, i, I, 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∼.

5. Conclusions

Over 490 M bytes of text was analysed and statistics of Polish phonemes, diphones and triphones were created. We do not claim that they are fully complete but the corpus was large enough, that they can be successfully used for language modelling. 27% of possible triples were detected as triphones, but some of them appeared very rarely. A majority of rare triphones came from foreign or twisted words. The full statistics are available on request by an email.

Acknowledgements

We would like to thank Institute of Linguistics, Adam Mickiewicz University for providing PolPhone - a software tool to make a phonetic transcription for Polish.

6. References

[1] E. Agirre, O. Ansa, D. Martínez, and E. Hovy, "Enriching wordnet concepts with topic signatures," *Procceedings of the SIGLEX Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, 2001.

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

[4] 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.

[5] B. Kollmeier and M. Wesselkamp, "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," *The Journal of the Acoustical Society of America*, vol. 102, pp. 2412–2421, 1997.

[6] 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.

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

[8] B. Ziółko, S. Manandhar, R. C. Wilson, and M. Ziółko, "Semantic modelling for speech recognition," *Proceedings of Speech Analysis, Synthesis and Recognition. Applications in Systems for Homeland Security, Piechowice, Poland*, 2008.

[9] 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.

[10] J. Holmes, I. Mattingley, and J. Shearme, "Speech synthesis by rule," *Language and Speech*, vol. 7, pp. 127–143, 1964.

[11] D. Ostaszewska and J. Tambor, *Fonetyka i fonologia współczesnego języka Polskiego (eng. Phonetics and phonology of modern Polish language)*. PWN, 2000.

[12] D. Oliver, *Polish Text to Speech Synthesis, MSc. Thesis in Speech and Language Processing*. Edinburgh: Edinburgh University, 1998.

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

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

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

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.