The use of statistics of Polish phonemes in speech recognition

Bartosz Ziółko[*], Jakub Gałka[**], Suresh Manandhar[*],
Richard C. Wilson[*], Mariusz Ziółko[**]

[*]Department of Computer Science, University of York,
Heslington, YO10 5DD, York, UK

[**]Department of Electronics, AGH University of Science and Technology,
al. Mickiewicza 30, 30-059 Kraków, Poland

[*]{bziolko,suresh,wilson}@cs.york.ac.uk ,
[**]{jgalka,ziolko}@agh.edu.pl

ABSTRACT

Statistical data on phonemes, useful in continuous speech recognition system, are presented. This paper explains basics of a simple system for phonemes, diphones and triphones statistics estimation from a text corpus of Polish language. Obtained results are presented for exemplar text database. Possible application of the statistics is suggested.

1. Introduction

Speech recognition (SR) seems to be currently in a dead end alley. Almost all solutions are based on the same general model [1]. Research is focused on improving it by adding additional elements. Such approach gives better results but it has to be accepted that there is a limit which cannot be overcome without changes in the general scheme. The method based on multi-level hidden Markov models (HMM) with features of constant-length frames found its usefulness in many applications. However, it seems not to be efficient enough to transcribe correctly any spoken language with large vocabulary. There are number of reasons. Some of them are very simple in their nature. The ASR system based on dictionary will never work properly for out-of-dictionary words. Grammar models will not deal properly with incorrect utterances while humans very often can. ASR systems try to recognize speech, while humans can also understand it and adapt to errors or unusual words. This causes the mentioned limit of classical

approaches. The typical SR method is in fact based on guessing and luck in a few of its steps. The speech is segmented into frames without any temporal speech motivated rules. HMM tries to find the closest transcription based on features what is in fact a kind of guessing. Such method works well enough for clearly spoken words with limited vocabulary. Noise, natural rate of speaking and not limited vocabulary cause many exceptions and information missing which HMM cannot deal with properly. Another very important problem is that people speak not as carefully as they write, while we expect a transcription produced by SR systems to be of the quality of our typed texts. It has to be accepted by both end users and researchers, that while we speak we do not always follow grammar rules and, what is more, errors in pronunciation give many exceptions independently of a dictionary size. This is why fitting a hypothesis using described language rules and dictionary does not always work.

The same problems occur in the case of names, out-of-language words, etc. SR systems try to fit the speech to language rules and static vocabulary what causes additional distortions. There is no simple solution for the described problems. We suggest use of collected phoneme statistics in a given language to use as, for example, a backup for a dictionary if there is a difficulty with matching features with any of the words from vocabulary.

The probability of neighboring of two phoneme classes is useful and important information which is not included in fitting speech features into possible words from a dictionary. Such a method gains on its simplicity and might be useful in many cases.

High-level speech modeling is a time-consuming process. In recognition applications, which are expected to work in real-time, it may cause unwanted delays and lead to further mistakes. Replacing the huge and complex language models with simplified solutions gives space for making system more flexible and susceptible for on-line modifications.

Statistics of signal properties are widely used in state-of-art speech recognition systems (HMM, GMM). Probability distributions are usually estimated with expectation-maximization iterative algorithm. Statistics of phonemes can be used as an initial condition for estimation procedure, and speed up estimation process, or can be used as a desired model itself.

2. Statistics extraction scheme

Obtaining of phonetic information from an orthographic text-data is not straightforward [2]. Transcription of text into phonetic data has to be applied first [3]. In this work, transcription algorithm bases on simple conditional rules and process that change sequences of letters into sequence of phonemes. Most important and phonetically significant co-articulation effects were taken into account as well [4]. This system performs in similar way as solutions used in typical speech synthesizer engines like Festival [5] or MBROLA [6]. Simplified (see Table 1) SAMPA [7] phonetic alphabet

was used for transcription output. Statistics can be now simply calculated by counting number of occurrences of each phoneme, phoneme pair, and phoneme triple in analyzed text.

Presented transcription method [see Table 2] is not correct according to linguistic rules and definitely not perfect but it was designed to find crucial rules of most common phoneme appearances and not to analyse rare ones or sophisticated pronunciation rules. The inaccuracy caused by errors in transcription is relatively low comparing with the inaccuracy caused by quality of a text corpus and the language nature itself. Our aim is to show high disproportions in different phonemes appearances. The application of presented ideas would need additional and time-consuming improvements. Statistics should be smoothed by assigning of higher probability values for phonemes rarely represented or absent in the corpus. It is very likely, that even in huge text-corpus, some phoneme combinations may never occur. It is a typical problem in statistical speech modeling. In this case, smoothing operation should be applied when ones use statistics as a statistical, generative model parameters (like HMM). The idea of such smoothing is described widely by companies collecting similar statistics and in literature on statistical modelling of languages [8]. Another issue is considering statistics of non-Polish phonemes in Polish. However it sounds ridiculous it has some reasons. A written text often contains single out-of-language words, mostly surnames and geographical names. There are very few non-Polish names which has a Polish transcription like Shakespeare – [pol. Szekspir]. Typically such names are used in original form and may cause some distortions in statistics.

3. Phoneme statistics of the Polish language

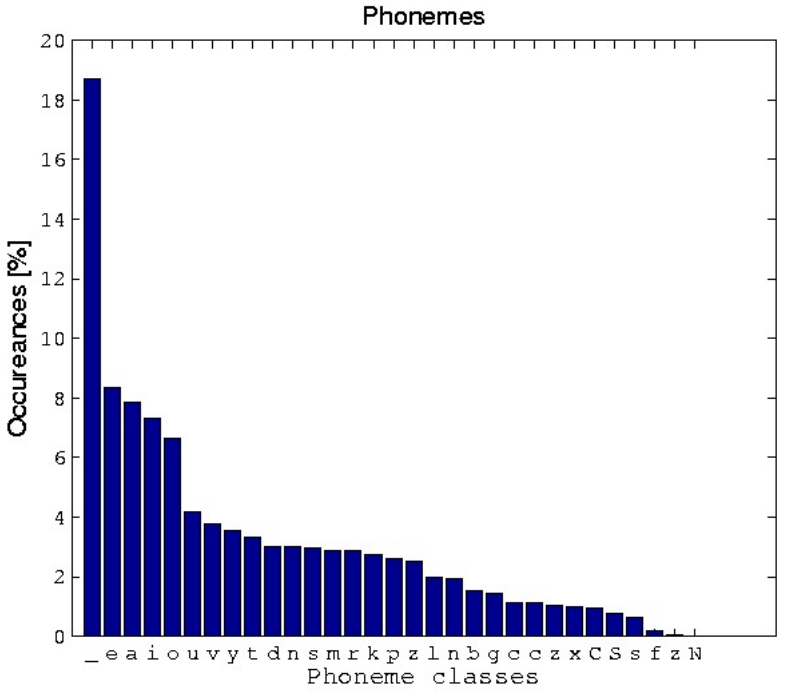
We analyzed the Polish language text corpuses to find information about phoneme statistics. Data for statistics were collected from literature and an MA thesis. 4466337 phonemes were analyzed and grouped finally in 31 classes as presented in Picture 1 and Table 3. A small number of phonemes, which should not exist in the Polish language, appeared as the result of analysis. It was probably caused by non-Polish words, names or phrases in analyzed text-corpus or errors in transcription. Those phonemes are not presented in the paper to simplify the statistics, as well as a few rare Polish phonemes are not included. 838 different diphones presented in Picture 2 were found for 961 possible combinations and 11052 different triphones (see Picture 3) for 29791 possibilities. This gives the conclusion that around 37 % of possible triphones are actually being used in the Polish language. The most popular triphones are presented in right side of Table 3.

Table 1. *Comparison of standard SAMPA notation and our modified SAMPA with examples in Polish*

SAMPA	example	transcription	SAMPA modification
i	PIT	pit	i
I	typ	tIp	I
e	test	test	e
a	pat	pat	a
o	pot	pot	o
u	puk	puk	u
e~	gęś	ge~s’	eu
o~	wąs	vo~s	ou
p	pik	pik	p
b	bit	bit	b
t	test	test	t
d	dym	dIm	d
k	kit	Kit	k
g	gen	gen	g
f	fan	fan	f
v	wilk	vilk	v
s	syk	sIk	s
z	zbir	zbir	z
S	szyk	SIk	S
Z	żyto	ZIto	Z
s’	świt	s`vit	ś
z’	źle	z’le	ź
x	hymn	xImn	x
ts	cyk	tsIk	c
dz	dzwoń	dzvon’	-
tS	czyn	tSIn	C
dZ	dżem	dZem	-
ts’	ćma	ts’ma	ć
dz’	dźwig	dz’vik	-
m	mysz	mIS	m
n	nasz	naS	n
n’	koń	kon’	ń
N	pęk	peNk	N
l	luk	luk	l
r	ryk	rIk	r
w	łyk	wIk	u
j	jak	jak	i

Table 2. *An example of transcription using our modified SAMPA*

Original text	Komputery umożliwiły tworzenie syntezatorów mowy dla praktycznych celów, dzięki czemu większe grono naukowców na świecie zainteresowało się zmienianiem tekstu na mowę.
SAMPA product	komputery umożliviuy tvożeńe syntezatoruv movy dla praktyCnyx celuv dzieuki Cemu vieukSe grono naukovcuv na świeće zainteresovauo sieu zmieńańem tekstu na moveu

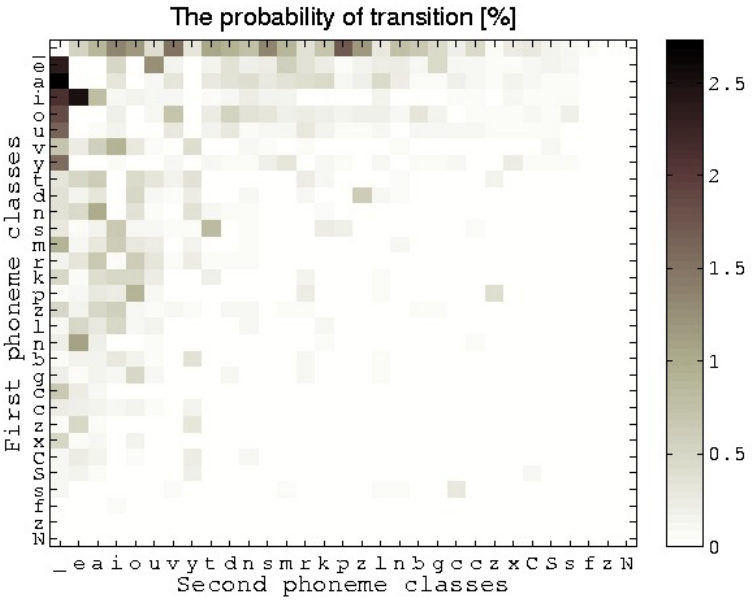


Picture 1. *Phonemes probability in the Polish language*

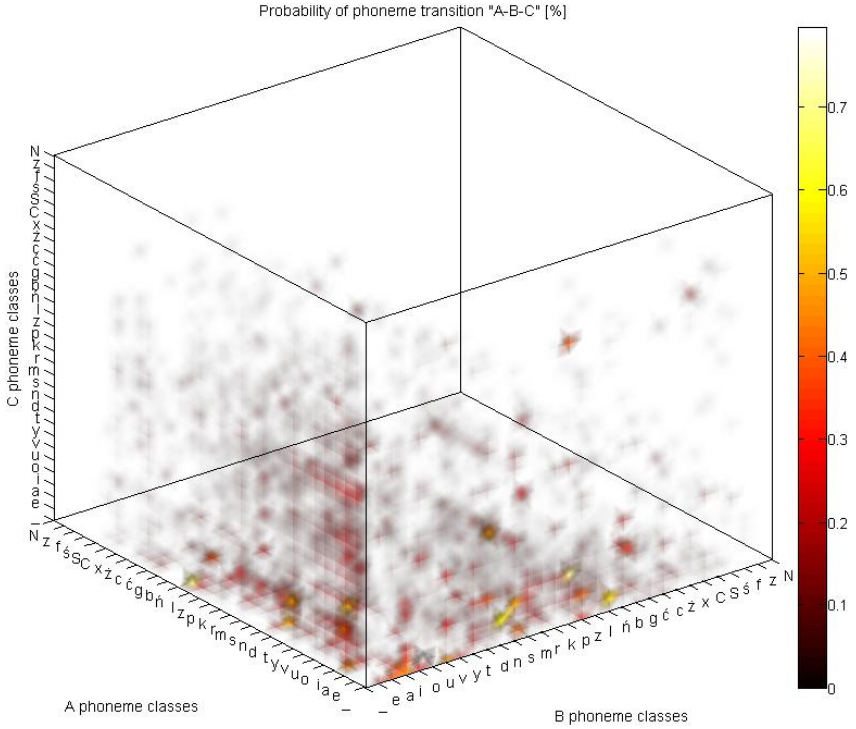
Table 3. *The probability of phonemes (left) and triphones (right) in Polish language*

Phoneme	Occurences	Percentage
_	836047	18.7189
e	372544	8.3412
a	350468	7.8469
i	326133	7.302
o	296398	6.6363
u	186239	4.1698
v	168853	3.7806
y	159448	3.57
t	149037	3.3369
d	134739	3.0168
n	134538	3.0123
s	131995	2.9553
m	128236	2.8712
r	127708	2.8593
k	121677	2.7243
p	115541	2.5869
z	113353	2.5379
l	88115	1.9729
ń	85346	1.9109
b	67565	1.5128
g	63332	1.418
ć	50748	1.1362
c	49835	1.1158
Z (ż)	45811	1.0257
x (h, ch)	44503	0.99641
C (cz)	42299	0.94706
S (sz)	33977	0.76074
ś	29047	0.65035
f	9120	0.20419
ź	2600	0.058213
N (ng)	1085	0.024293

Triphone	Occurences	Percentage
eu_	35249	0.7947
_po	31779	0.71646
ńe_	30741	0.69306
ieu	28510	0.64276
_na	26875	0.6059
_ńe	25104	0.56598
sie	22396	0.50492
na_	22057	0.49728
_si	21638	0.48783
_v_	20927	0.4718
dzi	20744	0.46768
vie	20630	0.46511
_i_	20485	0.46184
ie_	19977	0.45039
_za	18499	0.41706
em_	18410	0.41506
_a_	17591	0.39659
go_	17565	0.39601
_pZ	16629	0.3749
_do	16471	0.37134
_ie	16376	0.3692
ei_	16050	0.36185
ego	15003	0.33825
_vi	12686	0.28601
_o_	12608	0.28425
_z_	12558	0.28312
_vy	11688	0.26351
zie	11535	0.26006
pZe	11169	0.25181
_ia	11063	0.24942
by_	11046	0.24903
ia_	10941	0.24667
sta	10595	0.23887
ied	10574	0.23839
e_p	10493	0.23657
a_p	10475	0.23616
to_	10390	0.23424
ym_	10219	0.23039
ak_	10172	0.22933
va_	10117	0.22809



Picture 2. *Diphones probability in the Polish language*



Picture 3. *Tri-phones probability in the Polish language*

4. Backup of a dictionary

Typically, ASR systems performs recognition by finding most likely word sequence hypothesis

$$W^* = \arg\max_W P(W \mid X) = \arg\max_W \sum_{H,S} P(W) P(H \mid W) f(X, S \mid W), \quad (1)$$

with HMM-state sequence S and phonetic transcription H using dynamic programming algorithms [1].

Such a solution works pretty well assuming that spoken words are in the dictionary. Unfortunately it is a large simplification of the reality. Spoken language is quite irregular due to dialects, wrong pronunciation of non-native speakers or people with disabilities, names (especially geographical names which rarely exist in the dictionary) and the increasing process of mixing different languages (i.e. by introducing foreign names and product names). The last issue is a very important point in many non common languages like Polish. They adapt many words, especially English ones.

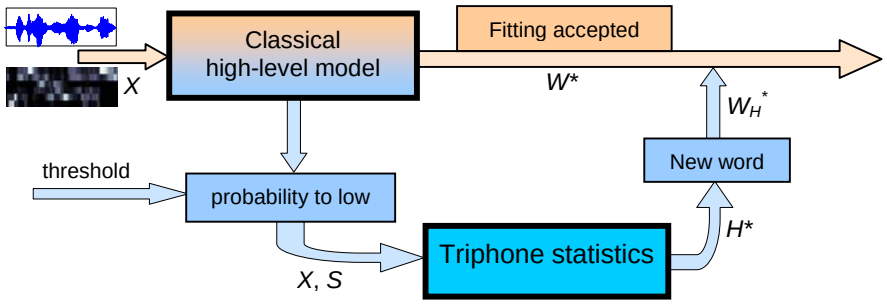
It is quite clear we cannot assume to have a dictionary with all words the speaker using ASR system may want to use. The system has to be prepared to deal with such exceptions. Currently the most common (if any) way is to force the user to spell a word (Dragon Naturally Speaking software). The system should at least try to present to the user a hypothesis of the non-dictionary word. The decision of spelling should be left to the user and not as the only way.

The possible solution is to use a triphone statistics as a backup of a dictionary as presented in Picture 4. The system would accept recognitions offered by HMM (with state sequence S) if the probability of their correctness exceeded some threshold. In case the hypothesis was not probable enough, another HMM using triphones statistics and original features X, would find the most likely phoneme sequence

$$H^* = \arg\max_H P(H \mid X) = \arg\max_H \sum_S P(H) f(X, S \mid H), \quad (2)$$

which may be transcribed into a proper corresponding word sequence W_H^*.

Adding to the problem with non-dictionary words described above, using dictionaries in SR systems for transcription has another disadvantage. A dictionary has to be extremely large to cover a whole language vocabulary. The amount of necessary operations is huge. This is why the use of a dictionary may be a bottleneck of the system. Unfortunately, current language modelling methods are probably not efficient enough to give us an opportunity to skip the use of dictionaries.



Picture 4. *Triphone statistics as a dictionary backup*

References

[1] Young, S. 1996. Large Vocabulary Continuous Speech Recognition: a Review, IEEE Signal Processing Magazine 13(5), 45-57.

[2] Holmes, J.N., Mattingley, I.G. & Shearme, J.N. 1964. *Speech synthesis by rule*, Language and Speech 7, 127-143.

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

[4] Ostaszewska, D. and Tambor, J. 2000. *Fonetyka i fonologia współczesnego języka polskiego* [in Polish]. PWN Warszawa.

[5] The Festival Speech Synthesis System, http://www.cstr.ed.ac.uk/projects/festival/

[6] The MBROLA Project, http://tcts.fpms.ac.be/synthesis/mbrola.html

[7] SAMPA – Computer Readable Phonetic Alphabet, http://www.phon.ucl.ac.uk/home/sampa/index.html

[8] Language Models in Speech Recognition, http://www.shlrc.mq.edu.au/masters/students/raltwarg/lmtoc.htm