Time Durations of Phonemes in the Polish Language

Bartosz Ziółko and Mariusz Ziółko

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

Abstract

Statistical phonetic data for Polish were collected. Phonemes are of different lengths, varying from 30 ms to 200 ms. Average phoneme durations are presented. A corpus of spoken Polish was used to collect such statistics from real language to apply it in an automatic speech recognition system. This natural phenomena could be used in speech parametrisation and modelling as an additional source of information for a case of speech segments analysis. The collected data are published in the paper, along with comments on the corpus and the method used. The obtained data were compared with the expected values according to phonetic literature.

Keywords: phoneme statistics, speech parameterisations, Polish

1. Introduction

The linguistic knowledge and statistics are important part of some engineering applications. Phoneme durations could be used effectively in speech modelling. There are segmentation and modelling methods which aim in locating phoneme boundaries in speech with unknown content (Glass, 2003; Ziółko et al., 2006). The segmentation combined with recognition could improve both processes, i.e. segmentation could be reset, if the recognised phoneme is of much different duration than the expected, statistical length. A phoneme duration can be seen as an additional parameter of a phoneme being recognised. The example of a word with phonemes of different durations is presented in Fig. 1.

2. Phoneme segmentation

Constant-time segmentation, i.e. framing, for example into 23.2 ms blocks (Young, 1996), is frequently used to divide the speech signal for processing. This method benefits from simplicity of implementation and results in an easy comparison of blocks, which are of the same duration. However, the uniform segmentation is perceptually unnatural, because the duration of phonemes varies significantly and is clearly longer than 23.2 ms.

Human phonetic categorisation is very poor for short segments (Morgan et al., 2005). Moreover, boundary effects provide additional distortions (partially reduced by applying the Hamming window), and such short segments create many more boundaries than there are between phonemes in the speech. The boundary effects can cause errors in speech recognition.

Additional difficulties appear because of the mixing of two phonemes in a single frame (Fig. 2). A smaller number of boundaries means a smaller number of errors due to the aforementioned effects. Constant segmentation therefore, while straightforward, risks losing valuable information about the phonemes due to the merging of different sounds into a single block. Moreover, the complexity of individual phonemes cannot be represented in short frames. The length of a phoneme can be also used as an additional parameter in speech recognition improving the accuracy of the whole process. The system has to know the expected duration of all phonemes to achieve larger efficiency.

Models based on processing information over long time ranges have already been introduced. The RASTA (RelAtive SpecTrAl) methodology (Hermansky and Morgan, 1994) is based on relative spectral analysis. The TRAPs (TempoRAl Patterns) approach (Morgan et al., 2005) is based on multilayer perceptrons with the temporal trajectory of logarithmic spectral energy as the input vector. It allows to generate class posterior probability estimates.

A number of approaches have been suggested (Stöber and Hess, 1998; Grayden and Scordilis, 1994; Weinstein et al., 1975; Zue, 1985; Toledano et al., 2003; Ziółko et al., 2006) to find phoneme boundaries from the time-varying speech signal properties. These approaches utilise features derived from acoustic knowledge of the phonemes. For example, the solution presented in (Grayden and Scordilis, 1994) analyses different spectra subbands in the signal. DWT (Discrete Wavelet Transform) was applied for phoneme segmentation task (Ziółko et al., 2006). Phoneme boundaries are extracted by comparing the percentage of signal power in different subbands. The Toledano et al. (Toledano et al., 2003) approach is based on spectral variation functions. Such methods need to be optimised for particular phoneme data and cannot be performed in isolation from phoneme recognition itself. ANN (Artificial Neural Networks) (Suh and Lee, 1996) have also been tested, but they require time consuming training.

Segmentation can be applied by the SM (Segment Models) (Ostendorf et al., 1996; Russell and Jackson, 2005) by searching paths through sequences of frames of different lengths. Such a solution means that segmentation and recognition are conducted at the same time and there is a set of possible observation lengths. In a general SM, the segmentation is associated with a likelihood and in fact describes the likelihood of a particular segmentation of an utterance. The SM for a given label is also characterised by a family of output densities which gives information about observation sequences of different lengths. These features of SM solution allow the location of boundaries only at several fixed positions which are dependent on framing (i.e. on an integer multiple of the frame length).

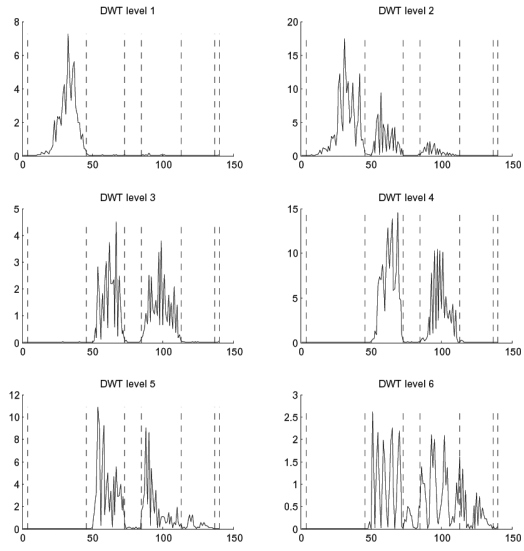

Figure 1: Discrete Wavelet Transform levels of word 'siedem' (Eng. seven) with segmentation (dashed, vertical lines) noted by a phonetician

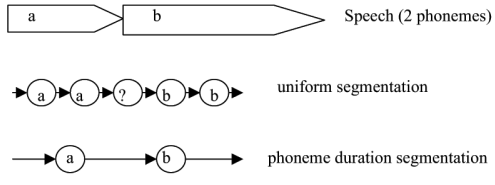

Figure 2: Comparison of the frames produced by constant segmentation and phoneme segmentation

The dynamic programming (Rabiner and Juang, 1993; Holmes, 2001) is a typical approach to phoneme segmentation for creating speech corpora. In speech segmentation it is used for time alignment of boundaries. The common practice is to provide a transcription done by professional phoneticians for one of the speakers in the given corpus. Then it is possible to automatically create phoneme segmentation of the same utterances for other speakers. This method is very accurate but demands transcription and hand segmentation to start with.

3. Experimental Data

The statistics were collected from CORPORA, created under supervision of Stefan Grocholewski in Institute of Computer Science, Poznań University of Technology in 1997 (Grocholewski, 1995). We investigated the male speakers only. Speech files in CORPORA were recorded with the sampling frequency $f_0 = 16$ kHz equivalent to sampling period $t_0 = 62.5$ μs.

The part of the database, which we used, contains 365 utterances (33 single letters, 10 digits, 200 names, 8 simple computer commands and 114 short sentences), each spoken by 28 adult males, giving 10220 utterances in total. One set was hand segmented. The rest were segmented by a dynamic programming algorithm which was trained on hand segmented one and based on transcriptions.

4. Statistics collection

The phoneme duration statistics were collected from MLF (Master Label Files) files attached to CORPORA. MLF is a standard solution, used in example in HTK (Hidden Markov Model Toolkit) (Young et al., 2005). They are defined in (Young, 1996) as index files holding pointers to the actual label files which can either be embedded in the same index file or stored anywhere else. The example of a part of an MLF from CORPORA is as follows:
"*/ao1m1ada.lab"
0 50000 sil
100000 1350000 a
1400000 1900000 d
1950000 3100000 a
3150000 4150000 m
4200000 4350000 sil
.
The description starts with name of an audio filename. Phone-me transcriptions are given (starting time, end time, a phoneme description) in following lines. The format ends with a dot. A basic time unit in this standard is 100 ns.

We summed all differences between starting and end times for all types of phonemes separately. The quantities of all types of phonemes in the corpus were also saved. Then the average phoneme duration was calculated, as the sum of durations divided by the number of phoneme occurrences. We calculated also standard deviation to evaluate how useful and trustworthy the data are.

5. Results

The statistics are presented in Tab. 1. The table includes silence and short pauses as they are a part of the standard notation. The average durations vary from 72 ms to 174 ms. Data in Tab. 1 include two units which are not phonemes. Abbreviation *sil* stands for silence and describes parts of audio files before and after speech. They are very short in this case, because CORPORA has very little silence in its records. Abbreviation *sp* stands for short pause. They appear in middles of words, for example, while speakers breath in.

CORPORA transcriptions are based on SAMPA notation with 37 symbols. Letters *ę* and *ą* are phonetically transcribed as *e_* and *a_* in CORPORA. However, these letters should be actually represented by two phonemes each. *ę* should be *e j∼* and *ą* should be *o w∼* but we are not able to detect these extra boundaries precisely enough. This is why we decided to keep them together as they are in the corpus. This is why *e_* and *a_* are the longest in Tab. 1. They represent two phonemes each, actually.

Table 1: Average duration of Polish phonemes with notations from CORPORA (Grocholewski, 1995) and SAMPA (Demenko et al., 2003)

CORPORA	SAMPA	av. duration [ms]	standard dev	example	transcr.
e_	e j∼	174	58	gęś	ge∼s'
a_	o w∼	166	52	ciąża	ts'ow∼Za
sz	S	152	59	szyk	SIk
s	s	132	46	syk	sIk
si	s'	130	45	świt	s'vit
c	ts	128	41	cyk	tsIk
a	a	127	48	pat	pat
ci	ts'	125	42	ćma	ts'ma
cz	tS	124	40	czyn	tSIn
f	f	122	64	fan	fan
zi	z'	115	33	źle	z'le
e	e	111	48	test	test
z	z	107	34	zbir	zbir
rz	Z	106	31	żyto	ZIto
drz	dz'	103	36	dźwig	dz'vik
o	o	103	35	pot	pot
h	x	100	45	hymn	xImn
dz	dz	100	35	dzwoń	dzvon'
u	u	99	42	puk	puk
t	t	98	52	test	test
dzi	dZ	98	27	dżem	dZem
k	k	94	45	kit	kitk
i	i	93	38	PIT	pit
p	p	93	41	pik	pik
n	n	91	41	nasz	naS
b	b	88	27	bit	bit
y	I	88	43	typ	tIp
m	m	86	34	mysz	mIS
d	d	83	29	dym	dIm
g	g	83	28	gen	gen
w	v	82	32	wilk	vilk
j	j	81	34	jak	jak
l_	w	79	33	łyk	wIk
ni	n'	76	33	koń	kon'
r	r	73	30	ryk	rIk
l	l	72	31	luk	luk
N	N	72	25	pęk	peNk
sp		68	28		
sil		15	26		

There are no numerical data like the presented in Tab. 1 available for Polish. However, we have found some information on this topic. According to (Wierzchowska, 1980), duration of phonemes is changeable and depends on speech ratio, type of utterance, localisation in a syllable and accents. What is quite constant is a ratio between durations of different phonemes. The longest ones are *e_* and *a_*. Then *a*, *o* and *e* are a bit shorter. Phonemes *i, y, u* follow them. *n* and *m* are average ones with *r* a bit shorter. *l* and *l_* are even shorter and *j* is the shortest one. It corresponds a bit to our results but not completly. Phonemes *e_* and *a_* are indeed the longest ones in both descriptions. We found that *a* and *e* are long, as described in (Wierzchowska, 1980) but *o* is average. We realised that *i, u* are expected to be quite long (Wierzchowska, 1980) but in CORPORA there are of average duration. Phoneme *y* is even shorter, however, (Wierzchowska, 1980) claims it should be long. Phonemes *n* and *m* are quite average as stated in (Wierzchowska, 1980). Phoneme *r* was found by us as a short phoneme what is in a contrast with (Wierzchowska, 1980). The experiment supports the opinion from (Wierzchowska, 1980) that *l*, *l_* and *j* are short. According to (Wierzchowska, 1980) there is a general rule that a duration is bigger for phones, for which a bigger number of parts of human speech system are necessary to be used.

The standard deviations are generally high. The ratio between average duration and standard deviation vary and is around 3 to 1. Phonemes *a, f, e, x, u, t, k, p, n, I, j, w*

have relativelly high standard deviations. It is probably a result of different ways of pronouncing these phonemes by different people.

There are some similar data in (Jassem, 1973). However, it does not present a complete list of durations. It gives some examples like that a transient can be up to 50 ms, *t* around 100 ms and *r* 20 ms. Again, some of these values corresponds to our results, like transient to short pause and phoneme *t*, but not all of them, like *r*, which is one of the shortest in our list, but its duration is 73 ms rather than 20 ms.

6. Conclusions

Phoneme statistical durations were presented for Polish. Their average values vary from 72 ms to 174 ms. It can be used efficiently in speech modelling and in automatic speech recognition systems. The obtained results are quite similar to what we expected after research in literature, however, no exact and complete numercial data of Polish phoneme durations were available till now. Standard deviations are usually around one third of average values which is high. It is caused by a fact that realisations of particular phoneme varies between different people. It is probably related to general rate of speaking.

7. Acknowledgements

This work was supported by MNISW grant OR00001905.

8. References

Demenko, G., M. Wypych, and E. Baranowska, 2003. Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis. *Speech and Language Technology, PTFon, Poznań*, 7(17).

Glass, J., 2003. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17:137–152.

Grayden, D. B. and M. S. Scordilis, 1994. Phonemic segmentation of fluent speech. *Proceedings of ICASSP, Adelaide*:73–76.

Grocholewski, S., 1995. Założenia akustycznej bazy danych dla języka polskiego na nośniku cd rom (Eng. Assumptions of acoustic database for Polish language). *Mat. I KK: Głosowa komunikacja człowiek-komputer, Wrocław*:177–180.

Hermansky, H. and N. Morgan, 1994. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589.

Holmes, J. N., 2001. *Speech Synthesis and Recognition*. London: Taylor and Francis.

Jassem, W., 1973. *Podstawy fonetyki akustycznej (Eng. Rudiments of acoustic phonetics)*. Państwowe Wydawnictwo Naukowe.

Morgan, N., Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard, and M. Athineos, 2005. Pushing the envelope - aside. *IEEE Signal Processing Magazine*, 22:81–88.

Ostendorf, M., V. V. Digalakis, and O. A. Kimball, 1996. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4:360–378.

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

Russell, M. and P. J. B. Jackson, 2005. A multiple-level linear/linear segmental HMM with a formant-based intermediate layer. *Computer Speech and Language*, 19:205–225.

Stöber, K. and W. Hess, 1998. Additional use of phoneme duration hypotheses in automatic speech segmentation. *Proceedings of ICSLP, Sydney*:1595–1598.

Suh, Y. and Y. Lee, 1996. Phoneme segmentation of continuous speech using multi-layer perceptron. In *Proceedings of ICSLP, Philadelphia*.

Toledano, D.T., L.A.H. Gómez, and L.V. Grande, 2003. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6):617–625.

Weinstein, C. J., S. S. McCandless, L. F. Mondshein, and V. W. Zue, 1975. A system for acoustic-phonetic analysis of continuous speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23:54–67.

Wierzchowska, B., 1980. *Fonetyka i fonologia języka polskiego (Eng. Fonetics and phonology of Polish)*. Zakład Narodowy im. Ossolińskich.

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

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

Ziółko, B., S. Manandhar, R. C. Wilson, and M. Ziółko, 2006. Wavelet method of speech segmentation. *Proceedings of 14th European Signal Processing Conference EUSIPCO, Florence*.

Zue, V. W., 1985. The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE*, 73:1602–1615.