Length of Phonemes in a Context of their Positions in Polish Sentences

Magdalena Igras, Bartosz Ziółko and Mariusz Ziółko
Department of Electronics, AGH University of Science and Technology,
al.Mickiewicza 30, 30-059 Kraków, Poland

Keywords: Phoneme Statistics, Spoken Language Processing, Polish.

Abstract: The paper presents statistical phonetic data of Polish collected from a corpus. Lengths of phonemes vary from 5 ms to 670 ms. Average durations of Polish phonemes are presented as well as an important anomaly of longer phonemes in the end of sentences, which is the main topic of the paper. This observation can be used in speech recognition for automatic insertation of dots and sentence modelling. Data of 45 speakers, 5130 sentences in total, were described and compared with the values taken from the phonetic literature.

1 INTRODUCTION

The linguistic knowledge and statistic parameters are important part of speech technology applications. Phoneme durations could be used effectively in speech modelling (Ziółko and Ziółko, 2011), for both, text to speech (Febrer et al., 1998) and speech to text systems (Linares et al., ; Pylkkönen and Kurimo, 2004). 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 its expected statistical length. Segmentation and acoustic modelling methods to locate phoneme boundaries in speech with unknown content (Glass, 2003), (Ziółko et al., 2011) were published.

Parameter of phoneme length is also often used in prosodic models for systems of automatic detection of punctuation and segmentation into sentences or topics (Shriberg et al., 2000; Kolář et al., 2004; Baron et al., 2002; Christensen et al., 2001) for other languages. Shriberg and Stolcke (Shriberg et al., 2000) use phone length normalization by phone specific values (means and standard deviations) and suggested also normalization of triphones duration.

It was observed that speakers usually tend to slow down their speech toward the ends of utterance units. Rate of speaking depends on individual characteristics of a person, but typically the last phonemes in a sentence are much longer (Demenko, 1999). Also for automatic punctuation annotation in Czech (Kolář et al., 2004) the phenomenon of preboundary lengthening is applied (only for vowels) described by following parameters: average duration of vowels, duration of the first and last vowel and duration of the longest and shortest vowel. The phenomenon was descirbed for Hungarian (Hockey and Fagyal, 1999) and Japanese (Shepherd, 2011) as well.

For Polish there are few published research on phonemes length modelling in a sentence. Some segmental features of vowels were investigated in logatomes (Frackowiak-Richter, 1973). Demenko (Demenko, 1999) described detailed analysis of correlations between place in a phrase and phonemes duration influenced by accent and intonational context. The research showed that in Polish last and one before the last vowel in a sentence is lenghtened regardless of accent presence, but the effect of lenghtening is stronger for accented syllables.

2 PHONEME SEGMENTATION

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

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 acoustic signals. The boundary effects can cause difficulties in

Igras M., Ziółko B. and Ziółko M..
Length of Phonemes in a Context of their Positions in Polish Sentences.
DOI: 10.5220/0004503500590064
In *Proceedings of the 10th International Conference on Signal Processing and Multimedia Applications and 10th International Conference on Wireless Information Networks and Systems* (SIGMAP-2013), pages 59-64
ISBN: 978-989-8565-74-7

speech recognition systems.

Additional difficulties appear when two phonemes are mixed in a single frame. Moreover, the contextual connections between neighbouring phonemes make frequency properties of the begining and the end of phonemes extremely irregular. A smaller number of boundaries means a smaller number of errors due to the effects described above. 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 phoneme durations can also help in locating ends of sentences. The system has to know the expected duration of all phonemes to achieve the greater efficiency.

A number of approaches have been suggested (Ziółko et al., 2011; Stöber and Hess, 1998; Grayden and Scordilis, 1994; Weinstein et al., 1975; Zue, 1985; Toledano et al., 2003) 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. Discrete wavelet transform was also applied for phoneme segmentation task (Ziółko et al., 2011). Phoneme boundaries are extracted by comparing the fractions of signal power in different subbands. Artificial neural networks (Suh and Lee, 1996) have also been tested, but they require long training. Segmentation can be applied by the segment models (Ostendorf et al., 1996; Russell and Jackson, 2005) by searching paths through sequences of frames of different lengths instead of using traditional hidden Markov models. 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.

3 EXPERIMENTAL DATA

The statistics were collected from CORPORA, created under supervision of Stefan Grocholewski in Institute of Computer Science, Poznań University of Technology (Grocholewski, 1997). Speech files in CORPORA were recorded with the sampling frequency $f_0 = 16$ kHz.

The part of the database, which we used, contains 114 short sentences, each spoken by 45 males, females and children giving 5130 utterances totally. Some part was hand segmented. The rest were segmented by a dynamic programming algorithm which was trained on hand segmented one and based on transcriptions and manually checked afterwards.

4 STATISTICS COLLECTION

The phoneme duration statistics were collected from MLF (Master Label Files) attached to CORPORA. MLF is a standard solution, used for example in HTK (Young et al., 2005). MLFs are defined as index files holding pointers to the actual label files which can either be embedded in the same index file or stored anywhere else (Young, 1996).

The description starts with name of an audio filename. Phoneme 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 0.1 μs. The example of a part of an MLF from CORPORA is as follows:

```
"*/jc1m1001.lab"
0 100000 sil
150000 850000 l
900000 1500000 u
1550000 2200000 b
2250000 2950000 i
3000000 3950000 si
4000000 5350000 cz
5400000 6400000 a
6450000 7150000 r
7200000 7850000 d
7900000 8550000 a
8600000 9600000 sz
9650000 10300000 o
10350000 10900000 w
10950000 11650000 y
11700000 12750000 p
12800000 13350000 l
13400000 15450000 a_
15500000 17850000 s
17900000 17950000 sil
.
```

5 PREBOUNDARY LENGHTENING

For analysis of the phenomenon of lenghtening phonemes in the end of sentence, we used 5130 sentences from CORPORA. Then each phoneme length was normalized by calculating ratio of phoneme length divided by mean length of the phoneme in the whole database. For each position i in each sentence we computed the mean value (μ) and standard deviation (σ) of the ratios dynamically, as the mean of ratios from position i=1 to the current i-th position. Then we verified, for which sentences the current value of μ, μ+ σ, μ+ 2σ and μ+ 3σ was exceeded by latest one, two, three or four phonemes in the sentence. We checked also for which sentence the thresholds were crossed before the end of the sentence.

Table 1: Average duration of Polish phonemes (in brackets divided by sum of all average durations) with notations from CORPORA (Grocholewski, 1997) and SAMPA (Demenko et al., 2003), standard deviations and ratio between deviation and average, No - number of appearance, μ - average[ms](%), dev - standard deviation.

CORPORA	SAMPA	No	μ[ms](%)	dev	$\frac{dev}{\mu}$	example	transcription
e_	e j∼	963	181 (4.46)	64	0.35	gęś	ge∼s’
a_	o w∼	2 293	171 (4.21)	59	0.35	ciąża	ts’ow∼Za
sz	S	27 49	155 (3.82)	63	0.41	szyk	SIk
s	s	4 251	134 (3.30)	48	0.36	syk	sIk
si	s’	2 117	131 (3.23)	49	0.37	świt	s‘vit
a	a	18 827	130 (3.20)	51	0.39	pat	pat
c	ts	2 514	129 (3.18)	44	0.34	cyk	tsIk
ci	ts’	1 880	127 (3.13)	45	0.35	ćma	ts’ma
cz	tS	1 670	126 (3.10)	43	0.34	czyn	tSIn
f	f	3 038	125 (3.08)	67	0.54	fan	fan
zi	z’	1 449	118 (2.91)	36	0.31	źle	z’le
e	e	13 034	113 (2.78)	51	0.45	test	test
drz	dz’	853	111 (2.73)	42	0.38	dżem	dZem
rz	Z	2 378	108 (2.66)	32	0.30	żyto	ZIto
z	z	2 489	108 (2.66)	35	0.32	zbir	zbir
dz	dz	509	105 (2.59)	30	0.29	dzwoń	dzvon’
o	o	10 844	105 (2.59)	37	0.35	pot	pot
h	x	2 383	102 (2.51)	46	0.45	hymn	xImn
t	t	5 874	102 (2.51)	56	0.55	test	test
u	u	5 444	101 (2.49)	46	0.46	puk	puk
dzi	dZ	1 229	100 (2.46)	29	0.29	dźwig	dz’vik
k	k	4 879	98 (2.41)	49	0.5	kit	kitk
i	i	7 106	95 (2.34)	40	0.42	PIT	pit
p	p	3 191	94 (2.32)	42	0.45	pik	pik
n	n	8 254	94 (2.32)	42	0.45	nasz	naS
y	I	5 533	90 (2.22)	46	0.51	typ	tIp
b	b	3 103	88 (2.17)	28	0.32	bit	bit
m	m	6 201	87 (2.14)	34	0.39	mysz	mIS
g	g	2 542	84 (2.07)	29	0.35	gen	gen
d	d	3 690	84 (2.07)	30	0.36	dym	dIm
N	N	279	83 (2.04)	25	0.30	pęk	peNk
w	v	3 973	83 (2.04)	31	0.37	wilk	vilk
j	j	7 067	83 (2.04)	35	0.42	jak	jak
l_	w	4 662	80 (1.97)	36	0.45	łyk	wIk
ni	n’	3 106	78 (1.92)	35	0.45	koń	kon’
r	r	6 657	76 (1.87)	35	0.46	ryk	rIk
l	l	5 707	74 (1.82)	33	0.45	luk	luk

5.1 Phonemes Duration

In total, 163 274 cases of Polish phonemes uttered by 45 speakers were analyzed. The statistics of phonemes durations are presented in Tab. 1, in descending order of the mean duration.

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

Then, we investigated probability distribution of the proportions of the duration of phonemes (example presented in Fig. 1) which appeared not to be Gaussian because of many much longer phonemes. The source of this anomaly is probably in longer phonemes in the end of sentences. The visual presentation of mean durations of phonemes from the whole database with their standard deviations are presented in Fig. 2.

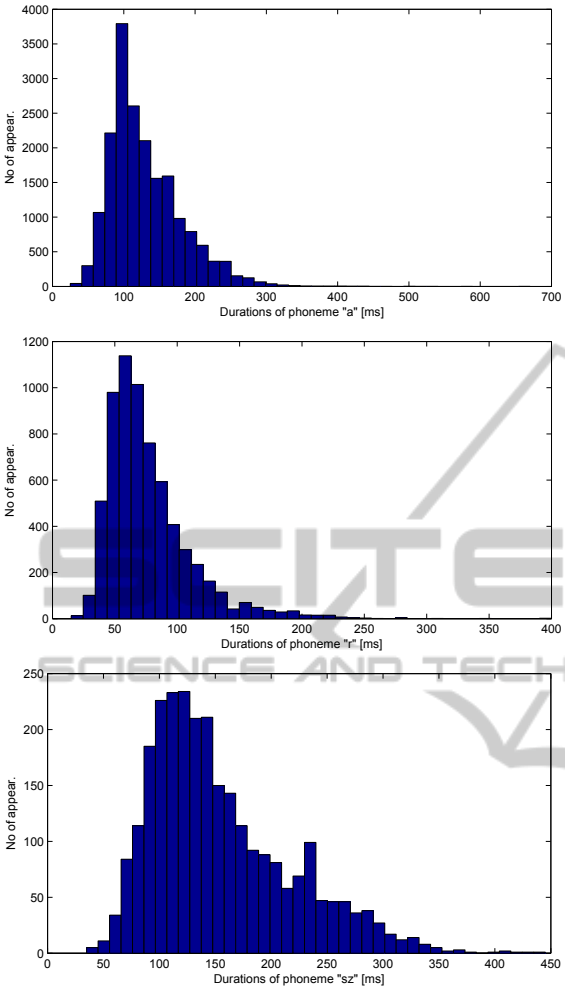

Figure 1: Probability distribution of the durations of example phonemes.

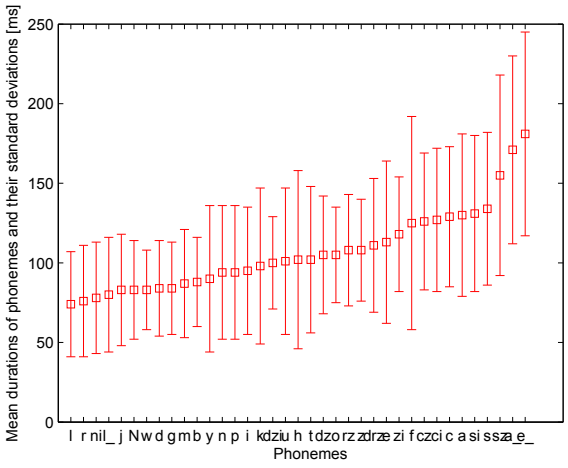

Figure 2: Mean durations and their standard deviations of phonemes from the whole database.

5.2 Phonemes Duration in Correlation with the Position in the Sentence

The weighted average of phoneme duration in the whole database was 104 ms. The average durations vary from 74 ms to 181 ms. Duration of phonemes is changeable and depends on speech ratio, type of utterance, localisation in a syllable and accents (Wierzchowska, 1980). A ratio between durations of different phonemes is quite constant. The longest ones are *e_* and *a_*. Then *a*, *o* and *e* are a bit shorter. Phonemes *i, y, u* follow them. Next, *n* and *m* are average ones with *r* a bit shorter. Phonemes *l* and *l_* are even shorter and *j* is the shortest one. A general rule is that a duration is bigger for phones, for which a bigger number of parts of vocal tract are necessary to be used (Wierzchowska, 1980).

It corresponds a bit to our results but not completely. 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 they 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.

The standard deviations of our results are generally high. The ratio between standard deviation and average duration vary and is between 0.29 and 0.68. Phonemes *t, f, y, k, r, u, e, h, l_, ni* (CORPORA notation) have relativelly high standard deviations. It is probably a result of different ways of pronouncing these phonemes by different people. The ratio of a standard deviation to an average value is lowest for phonemes *dz, dzi, rz, N, zi, b, z, c*.

There are some similar data in (Jassem, 1973). However, it does not present a complete list of phonemes durations. It gives some examples like that: a transient can be up to 50 ms, *t* around 100 ms and

Table 2: The percentage of phonemes much longer than the average for particular class according to its localisation in a sentence (b.f. - before last).

Location in sentences	Threshold $\mu+$			
	0	σ	2σ	3σ
The last phoneme	89.5	67.6	37.2	9.5
The one b. l.	65.6	27.8	2.9	0.0
The second b. l.	66.2	29.7	0.2	0.0
The third b. l.	60.3	22.6	0.0	0.0
Any other	97.8	65.9	2.5	0

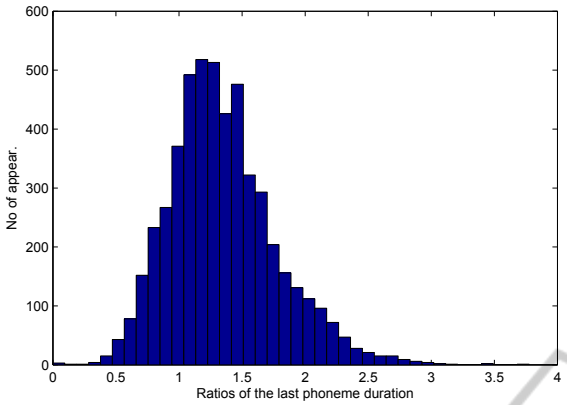

Figure 3: Probability distribution of the proportions of the last phoneme duration in the sentences to the mean phoneme durations.

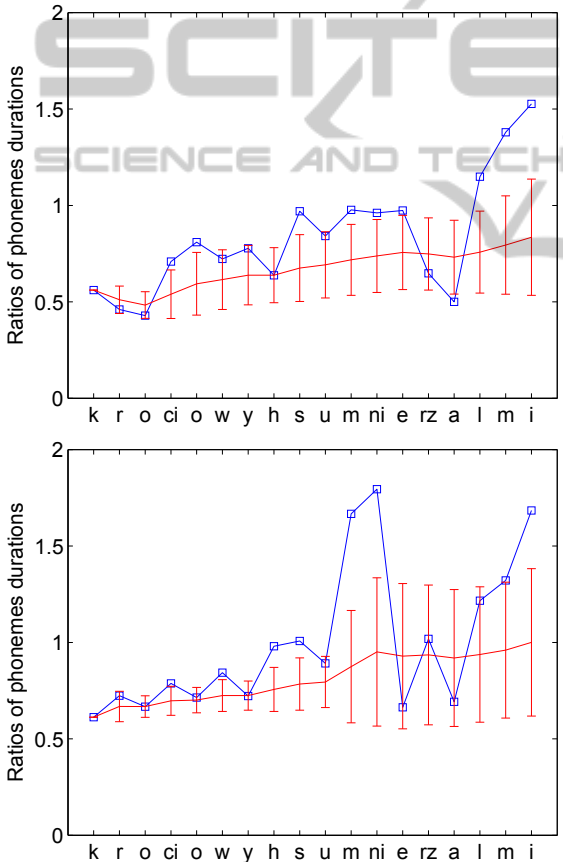

Figure 4: Example phoneme durations within a sentence (blue squares) with a dynamic average of the ratio (red line with standard deviations as column markers).

r usually 20 ms. Again, some of these values corresponds to our results, like transient to a 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 63 ms rather than 20 ms.

To ilustrate the tendency to lenghten phonemes in the end of a sentence, the percentage of phonemes much longer the average for particular class according to its localisation in a sentence was computed (see Table 2). It shows that for almost 90% of investigated sentences, the last phoneme durations exceeds the mean of antecedent phonemes ratios, while for majority of the sentences - it crossed also $\mu+\sigma$ threshold. Mean value of ratios of last phoneme duration in reference to the mean duration of the phoneme in the database was 1.35 (with standard deviation 0.42). Its probability distribution is presented in Figure 3. Several examples of sentences that illustrate the changes of the relative durations and their dynamic mean were presented in Fig. 4 and 5.

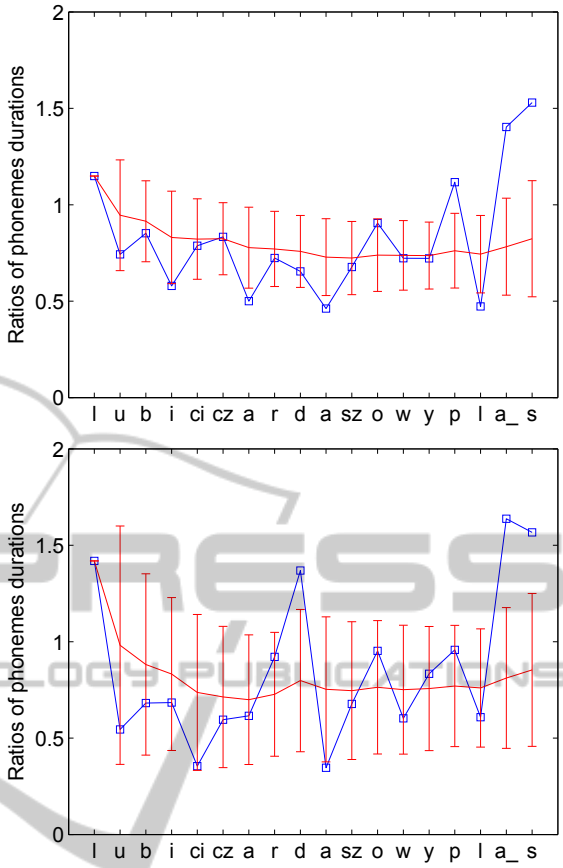

Figure 5: Example phoneme durations within a sentence (blue squares) with a dynamic average of the ratio (red line with standard deviations as column markers).

6 CONCLUSIONS

Statistics of phoneme durations in Polish were presented and compared with literature. The concept of applying them for sentence modelling was considered and evaluated. Around 37% of sentence ends

can be detected by analysis of phoneme length with only around 2.5% rate of false detections because phonemes in the sentence ends tend to be longer then other ones.

ACKNOWLEDGEMENTS

The project was funded by the National Science Centre allocated on the basis of a decision DEC-2011/03/D/ST6/00914.

REFERENCES

Baron, D., Shriberg, E., and Stolcke, A. (2002). Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. pages 949–952.

Christensen, H., Gotoh, Y., and Renals, S. (2001). Punctuation annotation using statistical prosody models. In *in Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, pages 35–40.

Demenko, G. (1999). *Analiza cech suprasegmentalnych języka polskiego na potrzeby technologii mowy [Eng. Analysis of Polish Suprasegmentals for Suprasegmentals for Speech Technology]*. Seria Językoznawstwo stosowane. Wyd. Naukowe Uniw. im. Adama Mickiewicza.

Demenko, G., Wypych, M., and Baranowska, E. (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).

Febrer, A., Padrell, J., and Bonafonte, A. (1998). Modeling phone duration: Application to catalan tts. In *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis. Jenolan Caves, Australia*, pages 43–46.

Frackowiak-Richter, L. (1973). *The duration of Polish vowels*. Speech analysis and Synthesis III, PWN.

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

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

Grocholewski, S. (1997). CORPORA - speech database for Polish diphones. *Proceedings of Eurospeech*.

Hockey, B. A. and Fagyal, Z. (1999). Phonemic length and pre-boundary lengthening: An experimental investigation on the use of durational cues in hungarian. *Proceedings of the XIVth International Congress of Phonetics Sciences, San Francisco*.

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

Kolář, J., Švec, J., and Psutka, J. (2004). Automatic punctuation annotation in czech broadcast news speech. pages 319–325, Saint-Petersburg. SPIIRAS.

Linares, G., Lecouteux, B., Matrouf, D., and Nocera, P. Phone duration models for fast broadcast news transcriptions.

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

Ostendorf, M., Digalakis, V. V., and Kimball, O. A. (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.

Pylkkönen, J. and Kurimo, M. (2004). Using phone durations in finnish large vocabulary continuous speech recognition.

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

Shepherd, M. (2011). The scope and effects of preboundary prosodic lengthening in Japanese. In *USC Working Papers in Linguistics*, pages 1–14.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics.

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

Suh, Y. and Lee, Y. (1996). Phoneme segmentation of continuous speech using multi-layer perceptron. In *Proceedings of ICSLP, Philadelphia*, pages 1297–1300.

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

Weinstein, C. J., McCandless, S. S., Mondshein, L. F., and Zue, V. W. (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. Phonetics and phonology of Polish)*. Zakład Narodowy im. Ossolińskich, Wrocław.

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

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

Ziółko, B., Manandhar, S., Wilson, R. C., and Ziółko, M. (2011). Phoneme segmentation based on wavelet spectra analysis. *Archives of Acoustics*, 36(1).

Ziółko, B. and Ziółko, M. (2011). Time durations of phonemes in polish language for speech and speaker recognition. *Lecture notes in artificial inteligence*, 6562:105–114.

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