DIFFERENT TYPES OF PAUSES AS A SOURCE OF INFORMATION FOR BIOMETRY

M. Igras, B. Ziółko

Department of Electronics, AGH University of Science and Technology
{migras,bziolko}@agh.edu.pl www.dsp.agh.edu.pl

Abstract: **Statistics of pauses appearing in Polish as a potential source for biometry information for automatic speaker recognition were described. The frequency of three main types of pauses (silent, filled and breath pauses) usage in monologues, as well as frequency of punctuation (commas and full stops) in their transcriptions were investigated quantitatively. Correlation between temporal structure of speech and syntax structure of the spoken language were examined statistically to verify usefulness of pauses detection for elaborating algorithms of automatic detection of punctuation for spoken Polish.**
Keywords : **pauses, fillers, punctuation, Polish**

I. INTRODUCTION

A set of common disfluencies interferes with sentences borders in spontaneous speech. The most important are: restarts, change of syntax during the utterance and inclusion of intervening sentences. Within words, the most frequent are repetitions, repairs and prolongations of conjunctives, prepositions and final syllables. As far as human perception can focus on the meaning of the utterance and extract the desired information, the automatic speech recognition system literally recognizes whole acoustic content of the speech signal [1]. As a result, the transcription is redundant with notation of disfluencies or slips of the tongue, but diminished of the other types of information present in signal, like punctuation. This information could be also used to differentiate speakers.

The research show three types of acoustic pauses in spoken language. The most intuitive is silence (s_p). Depending on the speaker and situational context, it may be characterized by different length.

Another type are filled pauses (f_p) - pseudo-words, that do not affect sentence meaning, like *yyy, eee, hmm, mmm, ym* (in SAMPA notation: *III*, *eee*, *xmm, mmm*, *Im*,), that perturb utterance fluency. They may often indicate need of insertion of comma or full stop in the adequate position in transcription. The sound of filled pauses are specific for language (in Polish the most common are *yyy/yh* and *mmm*, while for English - *um*) and specific for speaker's habits. The third sort of pauses that we consider are breath pauses (b_p) which strongly indicate insertion of the full stop in transcription.

Breath events [2] and filled pauses [3] can be automatically detected in a speech signal. It allows to apply this methods as a part of biometry systems in speaker recognition task.

Considering the origin of pause usage we marked out 1) regular natural pauses caused by respiration activity (breath pauses), 2) irregular intentional pauses, purposely used as a stylistic form, especially by professional speakers (silent pauses) and 3) irregular, unintentional disfluencies, effect of uncertainty, hesitation or short reflection, in speech of inexperienced speakers even 10-20 per minute (acoustic events like silent pauses or filled pauses).

Information on pauses is used in majority of algorithms of automatic punctuation detection [4], [5]. Some medical aspects of different types of pauses were investigated in context of affective state [6] and mental condition [7] of the speaker.

The obtained knowledge on pauses meaning can be merged with analysis of other temporal features (phoneme length, energy, fundamental frequency [8]) in order to build algorithms for punctuation detection in speech.

II. METHODS

The prepared corpus of spontaneous Polish speech consisted of different types of monologues in formal or half-formal situations. Total duration of recordings is 60 min, including utterances of 24 speakers (13 male, 11 female). Among them, there are both experienced or professional speakers (politicians, professors, professional translators) and inexperienced speakers (students).

The first group of recordings (30 min) are utterances from orations or public presentations: speeches and reports from European Parliament [9], sessions of faculty council, students lectures and reviews. All the speeches, although preceded by preparation of the speakers or supported by slides, were not read and are characterized by all the features typical for spontaneous speech. The second part of the corpus (30 min) consisted of recordings of real time translation of orations during European Parliament sessions [9]. The sort of utterances are specific kind of spontaneous speech, where the speech

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

rate of the translator is determined by the style of the speaker being translated. However, still they are situations of formularization of own utterance, which causes their spontaneous character and induces presence of imperfections characteristic for spontaneous speech.

For comparison with read speech, recordings from audiobooks and AGH Audio-Visual Speech Database [10] were used.

First we transcribed orthographically the content of the recordings to clean (skipping disfluencies, filled pauses or repairs) and syntactically correct texts. On the basis of the observation of the process, the factors affecting the imprecision and ambiguity of inserting punctuation in the transcripts were collected. One of the impediments was ambiguous intonation, especially in case of inexperienced speakers. It manifested by 'enumerating' tone of voice, which caused the same tone in commas and full stops or constructing multiple complex sentences with every clause starting with conjunctive pronounced with extended phonation. In such cases the decision of inserting comma or full stop remained subjective. When a speaker did not signalize the phrases and sentences border with their pronunciation, intonation or pauses, the punctuation was based on the meaning of the utterance. There also often occurred the bonding the last word of preceding sentence with the first in the next one. In translators group we usually observed specific disorder of phonotactics involving artificial prolongations of whole words. Transposals of functional elements of sentences and reorganization of the sentence were also frequent events. It is common for inexperienced speakers to place intervening sentences during the speech or abusing certain words like *let's say, just, simply* (language-specific conversational fillers/discourse makers).

For each transcription, the number of words, full stops and commas were counted. Then the statistics of sentences and phrases lengths were computed: mean length of a sentence and a phrase, as well as mean number of words in sentences and phrases. Then, in the places of punctuations signs, occurrences of pauses were verified. When a full stop were signalized by silent pauses, the time was tagged as s_p. (similarly for commas - s_p,), filled pauses - f_p. (commas - f_p.), b_p. for breath pauses (b_p, for commas). When no type of pause appeared, the place was tagged as n_p. (n_p,).

III. Results

Information of frequency of using punctuation signs in spoken language as phenomena determining speech rhythm were obtained by analyzing the quantity of full stops and commas in transcriptions. Fig. 1 shows meaning of the pauses in determining punctuation in speech. Fig. 2 presents the most frequent types of filled pauses. However, the usage of different types of pauses for signalization of punctuation is strongly individualized between speakers, as presented in Table 1.

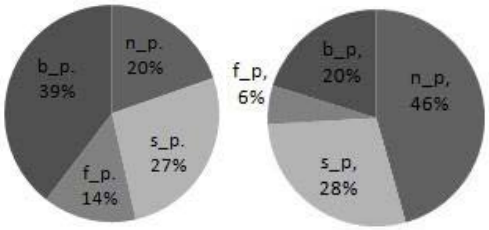

Fig. 1. Different types of pauses determining full stops and commas, and types of filled pauses signalizing punctuation

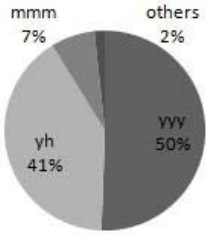

Fig. 2. Different types of filled pauses appearing in spontaneous speech

Table 1. Counts of pauses events denoting full stops and commas (P- results for presentations/orations, T- translation, C- entire corpus, * - lack of audible breaths in recording)

Fullstops

Rec.	#.	n_p.	s_p.	f_p.	b_p.
P1_5	17	2	11	4	*
P2_5	26	1	6	19	*
P3_3	15	1	14	0	*
P4_1	11	2	2	0	7
P5_2	10	4	0	2	4
P6_1	8	0	1	1	6
P7_1	4	0	1	1	2
P8_1	7	0	1	2	4
P9_1	6	0	0	4	2
P10_1	4	1	1	1	1
P11_5	29	1	13	0	15
P12_4	40	13	10	1	16
P	177	25	60	30	57
T1_1	7	4	1	0	2
T2_1	8	2	2	2	2
T3_1	8	3	4	0	1
T4_1	11	0	5	0	6
T5_1	8	4	1	0	3
T6_5	32	6	4	5	17
T7_5	18	12	1	1	4
T8_5	24	0	6	1	17
T9_2	9	0	3	1	5
T10_4	27	8	5	4	10
T11_1	5	2	0	0	3
T12_2	17	3	2	0	12
T	174	44	34	14	82
C	351	69	94	49	139

Commas

Rec.	#,	n_p,	s_p,	f_p,	b_p,
P1_5	71	27	35	13	*
P2_5	80	32	36	12	*
P3_3	24	9	15	0	*
P4_1	12	8	2	0	2
P5_2	22	14	2	0	6
P6_1	10	7	2	0	1
P7_1	14	6	3	1	4
P8_1	6	4	1	0	1
P9_1	11	7	2	2	0
P10_1	21	10	9	1	1
P11_5	51	6	11	0	34
P12_4	99	55	24	1	18
P	421	185	142	30	67
T1_1	7	5	2	0	0
T2_1	7	4	0	1	2
T3_1	4	3	1	0	0
T4_1	8	3	1	0	4
T5_1	8	5	0	1	2
T6_5	57	29	13	1	14
T7_5	41	20	11	1	9
T8_5	25	7	7	1	10
T9_2	13	5	1	0	7
T10_4	49	30	9	1	9
T11_1	13	6	2	0	5
T12_2	19	6	3	3	7
T	251	123	50	9	69
C	672	308	192	39	136

As we estimated, speech rate in spontaneous monologues is about 115 words per minute (with standard deviation between speakers is about 20 words/min). Mean length of sentence (containing average 19 words) was about 10 seconds, while mean length of a speech unit divided by punctuation (average 7 words) - 3.8 s. The results were similar for both orations/ presentations and real time translations.

Among all full stops in transcription, 39% are correlated with occurrences of breath pause, 27% silent pause, 20% filled pause. Among all commas, 28% are pointed by silent pause, 20% breath pause and 6% filled pause. Lack of any kind of pause (words bonding in pronunciation) was registered in 20% occurrences of full stop and 46% commas for spontaneous speech, and only for 1,3% full stops and 42% commas for read speech. Among all occurrences of filled pauses, 8% indicate full stops and 6% indicates commas, among breath pauses the proportions are, respectively, 10 and 11%.

The most commonly used types of filled pauses are: prolonged 'yyy' (a half of the cases), short 'yh' (41%) and 'mmm' (7% of counts). As for acoustically registered breath pauses, average for a speaker was about 11 breaths per minute. In normal physiological condition, at rest, the value of breath per minute is 12-20.

To investigate the influence of experience and oratorical abilities on pauses and speech rate, we divided corpus of spontaneous monologues into recordings of experienced speakers (professors and politicians) and inexperienced speakers (mainly students). Average values of selected temporal features of each group are compared in Table 2.

Table 2. Comparison of selected features for experienced and inexperienced speakers: average values and standard deviation (in brackets)

	Professional speakers	Inexperienced speakers
#words/minute	108 (23)	117 (26)
#words/sentence	17(4)	22(6)
#f_p/minute	4(3)	10(5)
n_p. [%]	12(15)	13(13)
s_p. [%]	26(31)	24(23)
f_p. [%]	10(12)	34(30)
b_p. [%]	50(17)	27(8)

As expected intuitively, professionals speak slowlier, with less disfluencies and formulate shorter sentences, which makes their speech more adjusted for efficient listening and understanding by recipients. Also their dynamic breathing rhytms are much more concordant with sentences boundaries (a half of fullstops were correlated with breath pauses). Such conscious dynamic breathing (taking a breath before beginning of a sentence or phrase) is one of the basic voice emission principles, often emphasized by authors of handbooks on speaking skills and techniques [11],[12].

IV. Discussion

While the full stops can be easily recognized by pauses detection, the commas does not seem to be possible to detect on the basis on pauses alone, without taking into account another parameters.

Both lack of punctuation and occurrence of disfluencies in spontaneous speech transcripts are factors that disturb their processing by natural language processing systems, parsers or information extraction systems, mainly because usually language models do not contain disfluencies and operate on full sentences [13]. Research on punctuation in spoken language can improve ASR systems, increase readability and usefulness of automatic transcripts for human, and adapt them to be processed by language models. Moreover, modeling of pauses in spoken language can be applied to more natural-sounding speech synthesis systems [14].

V. Conclusion

Beyond applications of the research on pauses for speech technology systems, it can be used also directly in the biomedical field.

Connotations between pauses and punctuation, as well as frequency and types of pauses vary between individuals and depend on speaking style of each person, speech quality, culture, experience and preparation for oral presentations. Thereby, the temporal features can be used for speaker biometry or evaluation of speaker oratorical skills.

Further research will cover also other reasons of pauses frequeancy and duration variability. One of them is a type of personality of the speaker or even mental illnesses - quantity and duration of silent pauses can be indicators of emotional state of the speaker or a measurable symptom of psychic disorders like schizophrenia or bipolar affective disorders. Frequency of filled pauses and breath pauses during monologues will be invesitgated as a significant marker of speaker stress and emotional arousal.

ACKNOWLEDGMENTS

The project was supported by The National Research and Development Centre granted by decision 072/R/ID1/2013/03.

REFERENCES

[1] M. Ziółko, J. Gałka, B. Ziółko, T. Jadczyk, D. Skurzok and M. Mąsior: *Automatic speech recognition system dedicated for Polish.* Proceedings of Interspeech, Florence (2011)

[2] M. Igras and B. Ziółko: *Wavelet method for breath detection in audio signals*, IEEE International Conference on Multimedia and Expo (ICME 2013) San Jose, California, USA July 15-19, 2013

[3] K. Barczewska and M. Igras: *Detection of disfluencies in speech signal*, Young scientists towards the challenges of modern technology : 7th international PhD students and young scientists conference : Warsaw, 17–20 September 2012, pp. 36

[4] D. Baron, E. Shriberg and A. Stolcke: *Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues.* (2002) 949-952

[5] E. Shriberg, A. Stolcke, D. Hakkani-Tur and G. Tur: *Prosody-based automatic segmentation of speech into sentences and topics* (2000)

[6] I. Homma and Y. Masaoka, *Breathing rhythms and emotions*., Experimental physiology, vol. 93, no. 9, pp. 1011–1021, Sept. 2008.

[7] V. Rapcana, S. Darcy, S. Yeap, N. Afzal, J. Thakore, and R.B. Reilly: *Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia*. Medical Engineering & Physics 32 (2010) 1074-1079

[8] M. Igras and B. Ziółko: *The influence of phoneme duration, energy and frequency features on the prominence of accent and sentence boundaries in spoken Polish*, Approaches to Phonology and Phonetics: APAP Lublin, 21-23.06.2013, Book of abstracts, pp. 28

[9] J. Loof, C. Gollan and H. Ney: *Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system.* Proceedings of Interspeech, Brighton (2009) pp. 88-91

[10] M. Igras, B. Ziółko and T. Jadczyk: *Audiovisual database of polish speech recordings*. Studia Informatica 33 2B (2012) 163-172

[11] M. Kotlarczyk, *Sztuka żywego sowa (eng. The art of living word)*, Gaudium Lublin, 2010.

[12] Z. Pawłowski et al., *Emisja głosu - struktura, funkcja, diagnostyka, pedagogizacja (eng. Emission of voice - structure, function, diagnostics, pedagogization)*, Wydawnictwo Salezjańskie Warszawa, 2008.

[13] E. Shriberg: *Spontaneous speech: How people really talk and why engineers should care*. In: in Proc. European Conf. on Speech Communication and Technology (Eurospeech. (2005) 1781-1784

[14] B. Zellner: *Pauses and the temporal structure of speech.* Fundamentals of speech synthesis and speech recognition (1994) 41-62