Proceedings of the Federated Conference on Computer Science and Information Systems pp. 165–168 DOI: 10.15439/2018F42 ISSN 2300-5963 ACSIS, Vol. 15

Automatic intonation-based keyword extraction from academic discourse

Iurii Lezhenin*, Vadim Diachkov *, Anton Lamtev *, Artyom Zhuikov*,
Natalia Bogach*, Elena Boitsova†, Evgeny Pyshkin‡
*Institute of Computer Science and Technology Peter the Great St. Petersburg Polytechnic University
194021 St. Petersburg Polytechnicheskaya, 21 Email: bogach@kspt.icc.spbstu.ru
†Institute of Humanities Peter the Great St. Petersburg Polytechnic University
194021 St. Petersburg Polytechnicheskaya, 19 Email: el-boitsova@yandex.ru
‡Software Engineering Lab. University of Aizu
Aizu-Wakamatsu, 965-8580, Japan Email: pyshe@u-aizu.ac.jp

***Abstract*—This paper examines the perspectives of intonation processing for automatic keyword extraction. Based on a discourse intonation model from D. Brazil, automatic tone pattern recognition in speech stream is performed. It is shown that automatic classification of tone patterns can be done using simple polynomials and correlation. The original software tool *PitchKeywordExtractor (PKE)* was applied to academic discourse (on-line lectures) to extract keywords. The results were compared to the output of popular tools for speech analytics: *VoiceBase* and *IBM Watson*. All the records were processed also with Praat software and annotated by human experts. Experiments show that none of the automatic systems outperforms the others and PKE, VoiceBase and IBM Watson have the identical error rates with respect to human expert opinion. It motivates further research and supports the tendency to integrate intonation and, more generally, prosody processing in automatic keyword extraction.**

I. Introduction

AUTOMATIC keyword extraction is an important operation of textual information processing, e.g., information retrieval, summarizing, indexing, etc. Speech content occupies a large share in the overall information environment being therefore a matter for automatic keyword extraction [1]. The common practice to retrieve the keywords from speech is limited to text-based supervised and unsupervised methods applied to automatic speech recognition (ASR) output. Meanwhile, speech has its inherent feature, namely, speech prosody, that can be processed automatically to leverage keyword extraction.

Prosody processing for keyword extraction has not been thoroughly studied so far. Nevetheless, during past decades, it was repeatedly highlighted that the involvement of prosody knowledge into speech processing frameworks contributes to their performance. Even though there exists a significant diversity in phonetic and phonological approaches to prosody modelling, it is widely acknowledged, that speech prosodic markers are stable. They can be directly measured and reliably classified by means of machine learning [2], [3], [4].

Speech prosody encompasses all suprasegmental speech phenomena, but the present research is focused on only one aspect of prosody, i.e., intonation, in terms of pitch or fundamental frequency F_0. This paper addresses speech intonation in context of automatic keyword extraction in English academic discourse and contributes to the approach presented in [5] towards better understanding of applicability of computational prosodic modelling for keyword extraction and possible benefits for existing speech keyword extraction techniques.

The rest of the paper is organized as follows: Section I establishes the research background; Section II describes automatic tone pattern recognition using polynomials; Section III outlines word-to-frame mapping; Section IV presents the results of polynomial model (p-model) accuracy evaluation and cross-validation of *PitchKeywordExtractor* [5] along with two popular speech processing tools, *VoiceBase* and *Watson*; Section V summarizes the paper.

Research background for this work originates from three areas: automatic keyword extraction techniques, integration of prosody knowledge into speech processing and automatic tone pattern recognition:

A. *Automatic keyword extraction techniques*

Automatic keyword extraction has been a subject of extensive and detailed research in the past. An extreme demand for fast, cost-effective and accurate keyword extraction algorithms is motivated by a growing amount of digital text information. Text mining, automatic data collection indexing, extractive and abstractive text summarization, keyword-based information retrieval as well as other related tasks and applications strongly rely upon the sets of keywords (e.g., [6]).

Detailed surveys of the state-of-the-art keyword extraction techniques can be found in [7], [8], [9]. A comparative analysis of automatic keyword extraction algorithms along with text summarization challenges was presented in [10]. Existing techniques can be classified by approach as supervised and unsupervised, the latter including simple statistic, linguistics, graph-based and hybrid. Supervised techniques require annotated training data, while unsupervised operate without preliminary annotation or labelling (e.g., [7]). A comprehensive study of performance for supervised ensemble methods and base learning algorithms (Naive Bayes, support vector machines, etc.) can be found in [11].

IEEE Catalog Number: CFP1885N-ART ©2018, PTI

Unsupervised methods were shown to be not less powerful than supervised ones; e.g., unsupervised morphology learning was found to produce similar results compared to a rule-based system [12]. In [9] automatic keyword extraction was performed very effectively with unsupervised graph-based keyword ranking. Keyword extraction from conversations using particle swarm optimization was shown to produce highly accurate query results [13].

B. Prosodic models in speech processing

Computational prosodic modeling integrated into speech processing workflow is a promising yet challenging area. Prosodic models have been reported to be helpful for various speech processing areas [14], e.g., automatic speech understanding (ASU), speech synthesis (TTS, text-to-speech) [15], [3], discourse tagging and segmentation [4] and automatic speaker verification [16]. It was shown that the combination of word and prosodic knowledge yielded the best results, with significant improvements over either knowledge source taken separately.

Prosodic models were found to increase speech recognition accuracy, having not been optimized for word recognition [4]. An impressive result in speech segmentation, where the prosodic model alone performed better than the language model alone [4], makes it reasonable to investigate the segmentation ability of prosodic models for keyword location within ASR output.

One of the key concepts of any prosodic model is a tone unit. In [17] the tone unit is defined as the realization of the information unit, which is extremely valuable in the context of keyword search. Both units are generated in the flow of discourse, referring to the phonological and grammatical levels respectively.

Prosodic models which motivated this research were Discourse Intonation model from D. Brazil et al. (communicative approach) [18] and Systemic Functional Linguistics of M. Hallidey et al. (grammatical approach) [17]. Both models operate with a set of tonal patterns, e.g., in Brazil model these are: *falling*, *rising*, *rising-falling*, *falling-rising* and *level*, each having a specific communicative payload. These tone patterns are connected to the categories of "given/new information" [17] or deemed to be "referring/proclaiming" [18], [19], [20], [21] This explicit relationship between intonation and meaning is exploited to search for keywords in speech.

C. Automatic tone pattern recognition

Location and classifying of tone units can be performed automatically. In [2] a 4-point model to approximate tone patterns is proposed and examined in contrast with other approximation models for tones (e.g., Bezier curves). [2] also presents a detailed study on 4-point model cascaded with several supervised classifiers and was shown to perform the best with a rule-based classifier. In [5] a continuous polynomial tone model (p-model) for Brazil tones was proposed. Functions inside p-model are used not to approximate pitch contours, but as ideal tone pattern sets to calculate correlations. Both models will be evaluated together to check p-model tone pattern recognition accuracy (see Experiment 1 in Section IV).

II. Automatic tone pattern recognition using polynomial model

Automatic tone pattern recognition implemented in PitchKeywordExtractor [5] is applied to locate a pitch pattern within a part of a record to retrieve a frame with a significant tone move. The task is to make a decision what pattern type is the closest to a frame of a record containing n readings of fundamental frequency (pitch) $F_0[k], 0 \leq k \leq n$ taken at the sample rate of f_d. $w_{min} \leq w \leq w_{max}$ is frame length range; $0 \leq l \leq n - w$ is frame shift from the first frame element. Thus, each frame contains lw-windowed signal $F_0[l : l + w]$ and one can easily see that these frames are of different length. To cope with it, a polynomial model can be easily scaled and shifted. Due to the pitch detection algorithm if for k-th sample $F_0[k]$ cannot be measured, it is defined as $F_0[k] = -1$. Median filtering is applied to smooth single pitch discontinuities.

A. Polynomial model (p-model)

We define 5 model functions ϕ_k, $k = 1..5$, which correspond to 5 Brazil tones - *falling*, *rising*, *rising-falling*, *falling-rising* and *level*. These functions are the 1st and 2nd order polynomials (Fig. 1).

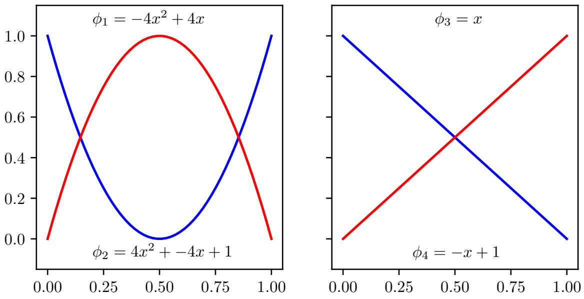

Fig. 1. p-model for Brazil tones

B. Classifier

To classify a frame by pattern type we evaluate its proximity to 4 model functions (except ϕ_5, "level"). ϕ_k, $k = 1..5$ define 5 decision regions separated by surfaces (Fig. 2).

Decision criterion to classify a frame to a region ϕ_i is

$$a_i = \frac{\sum_{k=l, F_0[k] \neq -1}^{w+l} (F_0[k] - \overline{F_0})(\phi_i((k-l)/w) - \overline{\phi_i})}{\sqrt{\sum_{k=l}^{w+l} (\phi_i((k-l)/w) - \overline{\phi_i})^2}},$$

and

$$\frac{a_i}{\sqrt{\sum_{k=l, F_0[k] \neq -1}^{w+l} (F_0[k] - \overline{F_0})}} = r_i \leq 1,$$

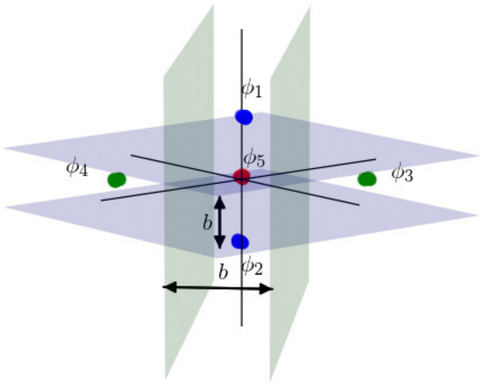

Fig. 2. Classifier decision regions $\phi_1 - \phi_4$. Parameter b sets the boundaries for *level* region ϕ_5.

where $\overline{F_0}$ is frame mean, $\overline{\phi_i}$ is model mean, r_i is normalized correlation; a_i is scaled correlation to distinguish any significant tone move from almost *level*. Decision criterion is scale and shift-invariant, thus, a_i are time and timbre-independent.

Decision is made as $k = argmax_i(a_i)$. If $max(a_k, b) \leq b$, where b is adjustable significance threshold and sets the *level* region, frame is classified as *level*, ϕ_5.

Thus, classifier outputs a pair (k, r). Frame overlaps are resolved [5] and the final frame set is successively transmitted to word-to-frame mapping.

III. WORD-TO-FRAME MAPPING

Frames where a tone move was detected during automatic tone pattern recognition and ASR output file are mapped to each other to locate a word within a frame. The goal of word-to-frame mapping is to find a word that was pronounced during a given interval defined by the frame boundaries; this word is deemed to be a keyword. Partial coincidence between segments and word timestamps is allowed and can be set as a parameter p.

Mapping in [5] could extract single words only. i. e. if one frame contained several matches with ASR output words, they were processed independently: e.g. for "computer science" the output list included both words one after another:"computer", "science". But very often a tone move refers to word collocation. Mapping is modified in order to extract keyphrases and word collocations as a single keywordlist entry. The following condition is checked sequentially for every frame:

If

$$\frac{L_{frame}}{t_2^{This_{word}} - t_1^{This_{word}}} \geq (1 + p), \tag{1}$$

then $[This_{word} \quad Next_{word}]$ is added to keyword list. This way keyphrases, e.g. of type "noun+noun" are constructed: "computer science", "artificial intelligence", "graduate student", etc.).

If

$$\frac{L_{frame}}{t_2^{This_{word}} - t_1^{Next_{word}}} \geq (1 + p),$$

then $[This_{word}, Next_{word}, Next\_next_{word}]$ is added to the list to produce constructions like "noun+preposition+noun" or "particle+verb+particle/attributive construction" (e.g. "place of interest", "to follow up", "to examine closely").

A further improvement of mapping may be achieved by *break* indices processing if ToBI annotated data are available.

IV. EXPERIMENTS

PitchKeywordExtractor implementation details, libraries and tools are described in [5]. New experiments are aimed at checking the applicability of proposed p-model in comparison with one of the best existing models (4-point model) and to disclose abilities of intonation-based keyword extraction to contribute to existing speech keyword extraction techniques.

Publicly available online lectures were used as samples of academic discourse to retrieve automatically pitch patterns (Experiment 1) and extract keywords (Experiment 2). Results on three speakers are shown in Table I, II:

Speaker 1 is Benjamin Elman from Harvard University's Fairbank Center for Chinese Studies The Great Reversal: The "Rise of Japan" and the "Fall of China" after 1895 as Historical Fables.

Speaker 2 is JoAnne Stubbe, MIT 5.07SC Biological Chemistry, MIT OpenCourseWare Lexicon of Biochemical Reactions: Cofactors Formed from Vitamin B12.

Speaker 3 is Patrick Winston, MIT 6.034 Artificial Intelligence, MIT OpenCourseWare, Introduction and Scope

All the records were processed with Praat software and annotated by human experts. *Expert* row in Table II is the absolute value of agreement between two human experts about tone patterns and keyword set for each *Speaker*.

A. Experiment 1. Pattern recognition

In Experiment 1 samples of academic speech were processed to check pattern recognition ability of p-model. p-model and 4-point model [2] are evaluated together to check p-model applicability for tone pattern recognition (Table I). Both models reveal almost identical recognition recall, calculated as

$$R = \frac{T_2}{T_1} 100\%,$$

where T_1 is a number of tones in total (tone units pointed out by *Expert*), T_2 is a number of tones found automatically.

TABLE I
PATTERN RECOGNITION WITH P-MODEL AND 4-POINT MODEL

Speaker	Tones in total	$A_{p-model}$	$A_{4-point}$
Speaker 1	50	52%	49%
Speaker 2	22	36%	36%
Speaker 3	40	22%	25%

B. Experiment 2. Intonation-based keyword extraction vs. other algorithms

Cross-validation of PitchKeywordExtractor (PKE) algorithm [5] vs. *Expert* and two popular speech processing tools, *VoiceBase* and *Watson* was performed. All the sets of

TABLE II
INTONATION-BASED KEYWORD EXTRACTION VS. HUMAN EXPERTS, VOICEBASE AND WATSON

Experiment	Speaker 1	Speaker 2	Speaker 3
E (human experts)	51	53	26
PKE	54	40	24
W(Watson)	58	51	18
VB(VoiceBase)	26	50	33
$PKE \cap E$	15	12	9
$W \cap E$	16	30	10
$VB \cap E$	10	11	4
$PKE \cap VB$	8	5	1
$PKE \cap W$	9	13	4
$VB \cap W$	16	9	3

keywords were compared and their intersections were counted. Numbers in cells show absolute value of keywords found. The observations that can be done based on Table II:

1) None of the systems outperforms the others
2) All the keyword sets found by the systems of automatic extraction (*PKE*, *W* and *VB*) have nearly the same intersection with *Expert*
3) All the keyword sets found by the systems of automatic extraction (*PKE*, *W* and *VB*) have nearly the same intersections with each other
4) There exist "core" keywords, extracted by either of the systems

V. CONCLUSION

Keywords are informative milestones of speech, therefore, they are frequently marked by prosodical emphasis; that is why specific discernible prosodic characteristics (tone moves) can mark keyword presence. Prosodic features in the form of F0 estimates allow computation of pitch contours along the utterances or single words, or over the length of windows positioned in a location of interest (e.g., around a word boundary). The algorithm is based on tone and information unit boundaries juxtaposition.

The goal of this paper is to provide evidence that automatic keyword extraction systems can benefit from intonation analysis. A software tool *PitchKeywordExtractor* was evaluated along with popular tools for speech analytics and revealed the identical ability to locate the keywords. A moderate percentage in intersections of human and automatic keyword sets, pointed out either by intonation-based and other algorithms, motivates further research towards the elaboration of a hybrid approach to automatic keyword extraction.

REFERENCES

[1] Polykarpos Meladianos, Antoine J-P Tixier, Giannis Nikolentzos, and Michalis Vazirgiannis, "Real-time keyword extraction from conversations," *EACL 2017*, p. 462, 2017.

[2] David O. Johnson and Okim Kang, "Automatic prosodic tone choice classification with brazil's intonation model," *International Journal of Speech Technology*, vol. 19, no. 1, pp. 95–109, Mar 2016.

[3] Anton Batliner and Bernd Möbius, *Prosodic Models, Automatic Speech Understanding, and Speech Synthesis: Towards the Common Ground?*, pp. 21–44, Springer Netherlands, Dordrecht, 2005.

[4] Elizabeth Shrieberg and Andreas Stolcke, "Prosody modeling for automatic speech recognition and understanding," 2002.

[5] Yurij Lezhenin, Artyom Zhuikov, Natalia Bogach, Elena Boitsova, and Evgeny Pyshkin, "Pitchkeywordextractor: Prosody-based automatic keyword extraction for speech content," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017, Prague, Czech Republic, September 3-6, 2017.*, 2017, pp. 265–269.

[6] Alan Darmasaputra Chowanda, Albert Richard Sanyoto, Derwin Suhartono, and Criscentia Jessica Setiadi, "Automatic debate text summarization in online debate forum," *Procedia Computer Science*, vol. 116, no. Supplement C, pp. 11 – 19, 2017, Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017).

[7] Slobodan Beliga, Ana Mestrovic, and Sanda Martincic-Ipsic, "Selectivity-based keyword extraction method.," *Int. J. Semantic Web Inf. Syst.*, vol. 12, no. 3, pp. 1–26, 2016.

[8] Slobodan Beliga, "Keyword extraction techniques," 2016.

[9] Yan Ying, Tan Qingping, Xie Qinzheng, Zeng Ping, and Li Panpan, "A graph-based approach of automatic keyphrase extraction," *Procedia Computer Science*, vol. 107, no. Supplement C, pp. 248 – 255, 2017, Advances in Information and Communication Technology: Proceedings of 7th International Congress of Information and Communication Technology (ICICT2017).

[10] Santosh Kumar Bharti and Korra Sathya Babu, "Automatic keyword extraction for text summarization: A survey," *CoRR*, vol. abs/1704.03242, 2017.

[11] Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, no. Supplement C, pp. 232 – 247, 2016.

[12] Yanzhang He, Brian Hutchinson, Peter Baumann, Mari Ostendorf, Eric Fosler-Lussier, and Janet B. Pierrehumbert, "Subword-based modeling for handling oov words inkeyword spotting," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7864–7868, 2014.

[13] D. Sowmya and J.I. Sheeba, "Keyword extraction using particle swarm optimization," *Procedia Computer Science*, vol. 85, no. Supplement C, pp. 183 – 189, 2016, International Conference on Computational Modelling and Security (CMS 2016).

[14] Janet Pierrehumbert, *Prosody, intonation, and speech technology*, p. 257–280, Studies in Natural Language Processing. Cambridge University Press, 1993.

[15] Grażyna Demenko, "Intonation processing for speech technology przetwarzanie intonacji na potrzeby technologii mowy," 2012.

[16] Mustafa Sönmez, Elizabeth Shriberg, Larry Heck, and Mitchel Weintraub, "Modeling dynamic prosodic variation for speaker verification," 01 1998.

[17] M. A. K. Halliday and William S. Greaves, *Intonation in the grammar of English / by M. A. K. Halliday and William S. Greaves*, Equinox Pub London ; Oakville, CT, 2008.

[18] David Brazil et al., *Discourse intonation and language teaching.*, ERIC, 1980.

[19] Miriam P. Germani and Lucia Rivas, "Discourse intonation and systemic functional phonology," *Colombian Applied Linguistics Journal*, vol. 13, no. 2, pp. 100–113, 2011.

[20] Dorothy M Chun, *Discourse Intonation in L2 – From Theory and Research to Practice*, 01 2002.

[21] Malcolm Coulthard and David Brazil, *The place of intonation in the description of interaction*, Linguistic Agency University of Trier, 1981.