9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Paragraph Prosodic Patterns to Enhance Text-to-Speech Naturalness

Àlex Peiró-Lilja, Mireia Farrús

TALN Research Group, Universitat Pompeu Fabra, Barcelona, Spain

alex.peiro@upf.edu, mireia.farrus@upf.com

Abstract

Speech synthesis has reached a reasonable high quality in recent years. However, there is still room for improvement in terms of naturalness and expressiveness when dealing with large multi-sentential discourse, since most text-to-speech synthesizers do not fully take into account the prosodic differences that have been observed in discourse units such as paragraphs. This work presents an implementation of paragraph-based prosodic patterns into the open-source MARYTTS platform, enriching its prosody output by means of intra- and inter-paragraph prosodic features. The set of characteristics include pitch decay, pitch range and speech rate variation (as intra-paragraph features), as well as paragraph break pauses and speech rate variation (as inter-paragraph features), previously analyzed in a large set of TED Talks and read-speech sections of the Spoken Wikipedia Corpus. The perception tests, performed both in English and German parametric voices, suggest that paragraph-based features should be further studied and taken into account on future implementations to synthesize large discourse speech.

Index Terms: parametric speech synthesis, text-to-speech, prosody, paragraph prosody patterns, MARYTTS.

1. Introduction

Over last years, great efforts have been made on improving intelligibility naturalness of text-to-speech (TTS) systems in both unit selection and parametric voices, leaving to the machine learning and statistical models a leading role on prosody enhancement [1, 2, 3, 4, 5, 6, 7, 8]. However, current state-of-the-art speech synthesis systems still lack of fully naturalness and expressiveness to resemble human speech, especially in multi-sentential discourse. Several works in literature have proved the existence of supra-segmental prosody patterns in discourse segments, e.g. the declination of speaker's pitch through both intra- and supra-sentential units [9, 10, 11], as well as in different speech styles [12], or depending on sentence position within the paragraph [13]. However, TTS systems fail in taking these supra-segmental characteristics into account to achieve a higher naturalness in the output speech.

In this paper, we present a first step towards the implementation of supra-sentential prosody feature patterns in a conventional multilingual TTS system, specifically at the paragraph level. In order to do so, we took the supra-sentential prosody patterns found in [13] and we incorporated them into the MARYTTS open-source platform. We implemented our modifications in a way that they complement the rule-based approach of the platform, thereby we can evaluate the resulting synthesis before pulling out the creation of new parametric models including these features at the input set. In addition, we carried out a study to extend the results found in [13] looking for prosody feature patterns such as pitch reset and pause between paragraphs, or the speaker rate variation along paragraphs.

The paper is structured as follows. Section 2 overviews implementations on speech synthetic voices to improve prosody in a sentential and supra-sentential approach. Section 3 provides a detailed description of the groups of paragraph prosody features studied and implemented in this work. Section 4 describes the open-source MARYTTS current system and our modifications. In Section 5 we show the evaluation results and discussion. And in Section 6 we present the conclusions.

2. Related Work

Improvement in naturalness and expressiveness in TTS systems has generally been attempted by increasing prosody variability based on inferred affective states [13]. In [3], a multiple linear regression model is presented to predict the most appropriate hidden Markov model (HMM) parametric voice style from a created set, and has also been implemented and evaluated including statistical models such as Gaussian mixture models (GMM) [2]. Moreover, the prediction of expressions from text and the synthesis of a particular expression have been integrated together [14]. In addition, prosody has been structured as a multi-level hierarchy for emotional speech synthesis [5], and its correlation with both hierarchical information structure and discourse has also been analyzed for speech synthesis purposes [15, 16, 17]. However, the general trend is to work on sentence level. Even current works based on deep learning techniques perform at sentence level as well to improve speech synthesis quality [4, 6, 8]; but there have been also attempts to work with prosody variations beyond the sentence, with the premise that a professional speaker modulates his voice from sentence to sentence [7]. Besides, different HMM topologies than conventional were trained using multiple-sentence short stories as utterances for sub-phonetic modeling in order to capture pronunciation variations [1].

The need of new databases to work beyond sentential segments has also been shown in the literature [18]. In this light, it seems there is especially a trend towards audio-book datasets as most of them contain expressive large multi-paragraph recordings for building voices [19] or working on prosody variation [3, 14, 7]. To the best of our knowledge, no previous work presents an implementation of paragraph-based prosody patterns as presented in this work, which uses a rule-based approach in a parametric TTS system.

3. Paragraph-based Prosodic Patterns

3.1. Intra-paragraph prosody feature patterns

This group of patterns is made up of the previous knowledge acquired in [13], which already found prosodic features with consistent patterns depending on the paragraph location at sentence unit level. In there, the authors analyzed a corpus of TED Talks consisting of more than 1300 talks with a large variety of speakers, finding out that some of the extracted features could differentiate, with statistical significance, first, middle and last sentence positions within a paragraph. Those features were related to F0 and intensity, pitch range, and speech rate. However,

10.21437/SpeechProsody.2018-124

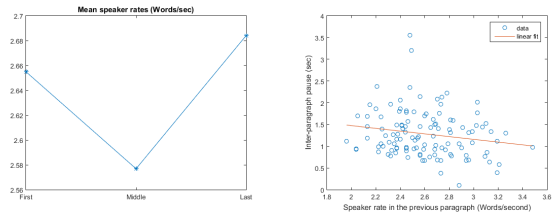
the findings were not implemented in a TTS system to validate them perceptually (see [13] for a more detailed explanation). Therefore, we included mean pitch, pitch range and speech rate in the intra-paragraph prosody features group, in order to generate more natural and, in turn, expressive speech. In this work, we assume all paragraphs have the same behavior.

3.2. Inter-paragraph prosody feature patterns

We define as inter-paragraph prosody features the ones that vary between two consecutive paragraphs or according to the paragraph location. For the inter-paragraph feature set, we carried out a study to investigate possible prosody patterns beyond intra-paragraph structure, such as pitch resets and pauses between consecutive paragraphs. For this purpose, we used the Spoken Wikipedia Corpus[1], a corpus of aligned spoken articles that includes English Wikipedia (SWC) from a diverse set of readers. We chose this corpus for the study because paragraph changes are very well defined. And although intra-paragraph-based patterns are taken from a more expressive corpus, we did not restrict the analysis to any speech style. Thus, we considered the SWC a good starting point to carry out a first analysis at inter-paragraph level.

We collected 40 multi-paragraph pieces of sections from different articles, obtaining a total of 117 paragraphs. Then we defined 3 different labels to categorize paragraphs: *first* (referred to the first paragraph of a text), *middle* (any paragraph between the first and the last one), and *last*. Therefore, paragraph-to-paragraph breaks are labeled as: first-middle, middle-middle and middle-last. We divided the study into three approaches: 1) *Local labeled features*: Computation of averages of pauses and pitch resets (pitch difference between first word of the paragraph and last word of the previous paragraph) of all paragraph breaks of each category, and the average speech rate (words/sec) of all paragraphs of each category. 2) *Local non-labeled features*: Without considering the paragraph break label, we computed the correlation between paragraph pause break duration and intra-paragraph features of the current (and also the following) paragraph. Also, the correlation between inter-paragraph pitch reset and intra-paragraph features of the current (and the following) paragraph was computed.

Although we could observe some pause and pitch reset patterns throughout the labeled paragraph breaks from approach 1, we only found significance in the speaker rate increase from middle paragraph to the last (pair-test, $p < 0.05$). From the approach 2, we found a significant correlation of -0.183 ($p < 0.05$) between paragraph speech rate and its pause break duration at the end. These statistically significant patterns (see Figure 1) were also implemented together with the intra-paragraph prosodic patterns.



(a) *Speech rate pattern along paragraphs* (b) *Speech rate and inter-paragraph pause correlation*

Figure 1: *Inter-paragraph significant patterns.*

4. Implementation in MARYTTS

4.1. MARYTTS system overview

MARYTTS[2] is a multilingual client-server TTS platform with an open-source license, purely written in Java. It stands out (among others) in the ease of code accessibility and good organization of the TTS modules and their corresponding classes. To our purposes this was a clear advantage, as we needed to understand the architecture before implementing our modifications. Moreover, although the project was originally designed for the German language [20], MARYTTS is currently a multi-language TTS that includes English, the language used for our previous analyses. The system has an internal representation language based on XML format, called MaryXML. So, all TTS modules contribute to the creation of the MaryXML file by adding on it the corresponding elements that give way to the synthetic voice.

Current MARYTTS' way to treat a multi-paragraph input plain text is by splitting up its paragraphs (delimited by line breaks) at first when the server gets the request. Then, these paragraphs are processed one by one as independent input texts. At the end, the generated MaryXML file marks up the limits of each paragraph, which shows up the systems' awareness of multi-paragraph texts, but it is not fully reflected in the predicted prosody. A linear decay function implemented in the prosody module inserts a mean and range pitch attributes at every phrase in the MaryXML file in order to simulate the pitch and range decay along the paragraph. However, the pitch range set does not make any effect on the output, at least, when using parametric voices because there is no code to process it; thus, the pitch contour comes out quite flat. Besides, the boundary pause duration model only predicts a 400 ms duration for any boundary element. Our modifications and new implementations on MARYTTS have been performed preserving the original structure and logic of the system. The bold modules shown in Figure 2 are the ones that we modified, which are detailed in the following subsection.

4.2. Prosodic modifications

4.2.1. Prosody module

In the prosody module of MARYTTS, accents, boundary positions, tone and break indexes are set sentence by sentence in a rule-based approach using ToBI or GToBI (a German adaptation of ToBI) annotation, and they are added in the MaryXML file afterwards. These rules are located in another XML file, and each language has its own with their particular rules. Boundaries are set depending on the type of sentence and its location.

We have extended the boundary rules set to include more break index (bi) levels for paragraph breaks according to their position in the text: first-to-middle (bi=6), middle-to-middle (bi=7), middle-to-last (bi=8) paragraph, and end-of-text (bi=9). We also adapted the code to check which paragraph is being processed, thus, assigning to it the corresponding paragraph break index. Only for declarative end-of-paragraph sentences boundary rules were added to the English and German XML files.

4.2.2. Acoustic parameters module

Once, having the new break indexes, we could enrich the boundary model for pause duration prediction, which is required at this stage. These durations were set depending on the break in-

[1] http://nats.gitlab.io/swc/
[2] http://mary.dfki.de/

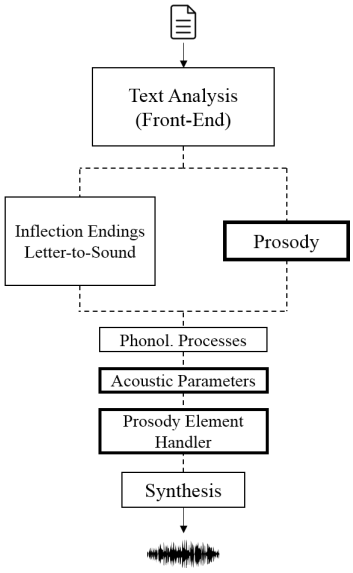


Figure 2: *MARYTTS basic architecture. Highlighted modules are those that we have modified.*

dex level: 290 ms (bi=4), 490 ms (bi=5), 1230 ms (bi=6,7,8), and 300 ms (bi=9). First- and middle-paragraph break pause values were taken from the polynomial fitting line (Figure 1b) as a function of the synthetic voice average speech rate (in words/second) of the English parametric voice. The rest of pause values were set empirically.

4.2.3. Prosody element handler module

This module is a prosody post-processing before synthesis. Here we added two functions that follow the corresponding pattern behaviors of pitch range (Fig. 4a) and speech rate (Fig. 4b) found in [13], setting the parameter proportional values empirically. Both functions were implemented in a rule-based approach depending on the sentence location in the paragraph: *first* (first sentence), *middle* (any sentence between the first and the last), and *last* (last sentence). We considered pitch range as the difference between first and last F0 values of a sentence, then by controlling the F0 contour slope we could modify this difference. Speaker rate is adjusted by increasing or decreasing predicted phoneme durations proportionally. Moreover, the number of words were also taken into account to provide a certain bias over sentence length variation. The significant pattern shown in Figure 1a is also implemented by a factor parameter depending on the location of the paragraph.

5. Evaluation and Discussion

We designed a perception test to evaluate the implementations both in English and German languages. Although the presented analyses were performed in English, the modules we modified are generally used for all languages currently supported by the system, and only the new specific prosodic boundary rules for declarative German sentences had to be added. German was also evaluated in a twofold objective: (i) to see whether our implementations based on previous English analyses are suitable for other languages, (ii) to integrate it in the KRISTINA project, which uses a German MARYTTS parametric voice for its intelligent embodied conversation agent, being German one of the main supported languages [21].

Since multi-paragraph perception tests can be difficult to assess due to samples length we selected for both languages no more than three texts. Two of them consisted of three paragraphs, and a third one of two paragraphs. English texts were taken from sections of Spoken Wikipedia Corpus articles not used in the inter-paragraph prosody features study, and the German ones from on-line local newspaper articles. Because the changes were not intended to produce a huge difference compared with the baseline synthesis, we chose samples with short paragraphs (3-4 sentences the longest paragraphs) to prevent subjects losing the objective of the evaluation. In total, 27 subjects took part in the experiment, and they were asked not to focus neither on intelligibility nor quality rather on the naturalness (*How close the speaker is from human natural voice reading aloud*) before starting.

The perception test consisted of two parts. First part was a Mean Opinion Score (MOS) for intra-paragraph prosody patterns evaluation. Only the largest paragraphs of each sample text were generated with both baseline system (BS) and baseline together with intra-paragraph prosody patterns (BS+IntraPP), thus, in total two versions of three paragraphs. The subjects evaluated in a 1-5 Likert scale (being 1 the worst and 5 the best) each version of each paragraph without the knowledge of what was which. The second part of the experiment was a pairwise test, which forced the subjects to decide between BS with or without adding together intra- and inter-paragraph prosody patterns (BS+IIPP). This time, the samples were the complete texts. The reason why we did not include inter-paragraph in the MOS was because we wanted the subjects to focus on the differences between baseline and modified synthetic voice with the intra-paragraph features. Otherwise they would probably have decided based on the inter-paragraph pauses, which is the most notorious feature. The parametric voices chosen to generate the samples are both feminine: *cmu-slt-hsmm en US female hmm* (English US) and *bits1-hsmm de female hmm* (German). The paragraph and complete text samples were about 30 seconds and 50 seconds long, respectively. The texts were shown next to the audio clips in the two tests.

Table 1: *MOS results in English*

	English	
	BS	BS+Intra-PP
P1	2.48 (0.68)	3.14 (0.91)
P2	2.71 (1.01)	3.29 (1.01)
P3	3.14 (0.85)	2.86 (1.15)
Avg Score	**2.79 (0.34)**	**3.10 (0.22)**

Table 2: *MOS results in German*

	German	
	BS	BS+Intra-PP
P1	2.50 (0.55)	3.17 (0.75)
P2	3.00 (1.10)	3.17 (0.75)
P3	3.17 (0.40)	3.33 (0.82)
Avg Score	**2.89 (0.35)**	**3.22 (0.10)**

The overall results of the MOS test tell that the subjects preferred our modified version over the baseline with an 11% of relative difference in both languages (See Tables 1 and 2). We perceived the German parametric voice was richer in prosody and sounded smoother than the English, and the total baseline scores are consistent with such perception, showing a difference of 3.6% between German and English. Baseline got a better score only in the third paragraph of English voice evaluation,

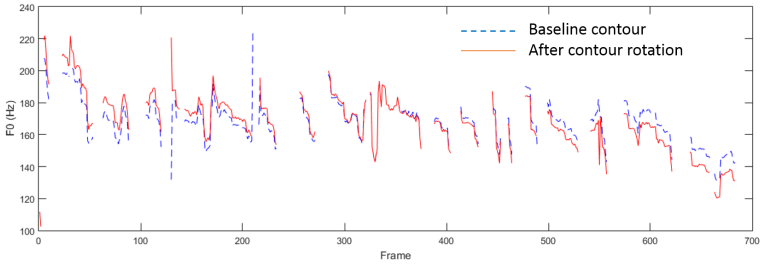
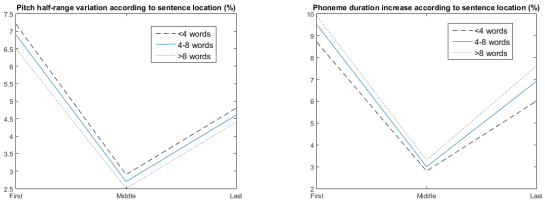
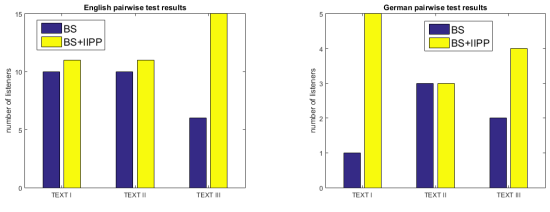


Figure 3: *A comparison between baseline generated F0 contour of an English sentence and after accentuation of the contour slope.*



(a) *Pitch half-range* (b) *Increase of phoneme dur.*

Figure 4: *Feature proportion variation according to paragraph location.*



(a) *English texts* (b) *German texts*

Figure 5: *Pairwise test results.*

which was the longest. Also the deviations are generally high, probably due to subjects did not have any previous reference to compare with the test samples, so the scores vary a lot between each subject in each sample. The pairwise test results also show a preference to our version samples. The third text BS+IIPP generated version was the most agreed between the English test subjects, which contained the third paragraph that was low-scored in the MOS test. As expected, BS+IIPP version of German texts were chosen with a higher agreement than the English. Only the second text got a tie. 59% of subjects voted the BS+IIPP version in English, and 67% the BS+IIPP version in German.

In summary, the results suggest an improvement over the baseline MARYTTS parametric English and German voices. But because our implementations have been performed at the rule-based post-processing level, the variability of the presented prosody feature patterns cannot be fully represented. That is why perhaps there were some disagreements, such as the mentioned case of the third English text, in which its inter-paragraph pause was probably more adequate for the subjects after a long paragraph. Or in the contrary, for the second German text, which both versions got the same votes, the paragraphs were short and they may needed shorter pauses. Moreover, intra-paragraph feature patterns may vary depending on the specific text content or length, and we implemented a generic behavior for any case, which could even negatively affect the naturalness of the voice. For instance, although we followed a specific pattern found in a previous work to set a degree of pitch range and declination along a sentence, it might not be the same in each sentence location. Probably boundary pauses and pitch declination were the two features that stand out the most perceptually. However, these results will allow us to keep this line of research to find out more suitable ways on integrating supra-sentential patterns in a TTS system.

6. Conclusions

We have presented an implementation of paragraph-based prosody feature patterns in a conventional statistical parametric TTS with the aim to improve its naturalness. In particular, we took the results of a previous study [13], where the authors found significant prosody patterns related to paragraph sentence location when analyzing a large speech corpus of TED Talks. These features are: speech rate, mean pitch, and pitch range, which we have grouped and called intra-paragraph-based prosody feature patterns. Moreover, we extended this idea and we carried out a study using the Spoken Wikipedia Corpora. We could find patterns with statistical significance, such as pause duration between paragraphs depending on the speech rate, and speech rate difference from middle paragraphs to last. So, we created another group: the inter-paragraph prosody patterns. Both groups have been implemented without modifying the MARYTTS architecture.

In general, the results of MOS and pairwise tests show a preference to our system version both in English as in German parametric voices. We found some limitations trying to implement these prosody patterns in a rule-based and post-processing approach, as there we could not cover the whole variability of the problem. However, based on the presented results and on the fact that we could find prosody patterns at inter-paragraph level, we are fully encouraged to continue in this research line. We aim to create parametric voices with complex statistical models that include new input features to deal with the presented prosody patterns.

7. Acknowledgements

The authors would like to thank all listeners that took part in the experiments. This work is part of the KRISTINA project, which has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under the Grant Agreement number H2020-RIA-645012. The second author is partially funded by the Spanish Ministry of Economy, Industry and Competitiveness through the *Ramón y Cajal* program.

8. References

[1] K. Prahallad, A. W. Black, and R. Mosur, "Sub-Phonetic Modeling For Capturing Pronunciation Variations For Conversational Speech Synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2006.

[2] O. Türk and M. Schröder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 965–973, 2010.

[3] M. Charfuelan, "MARY TTS HMM-based voices for the Blizzard Challenge 2012," in *Proceedings of Blizzard Challenge 2012. SynSIG Blizzard Challenge, September 14, Portland, Oregon, USA*, 2012.

[4] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7962–7966.

[5] Z. Wang and Y. Yu, "Multi-level prosody and spectrum conversion for emotional speech synthesis," in *Proceedings of the International Conference on Signal Processing (ICSP)*, vol. 2015-January, no. October, 2014, pp. 588–593.

[6] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based Recurrent Neural Networks," in *Proceedings of the INTERSPEECH*, 2014, pp. 1964–1968.

[7] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 2217–2221.

[8] H. Liu, H. Lu, X. Shao, and Y. Xu, "Model-based parametric prosody synthesis with deep neural network," in *Proceedings of the INTERSPEECH*, 2016, pp. 2313–2317.

[9] J. Kreiman, "Perception of sentence and paragraph boundaries in natural conversation," *Journal of Phonetics*, vol. 10, pp. 163–175, 04 1982.

[10] J. Hirschberg and J. Pierrehumbert, "The intonational structuring of discourse," in *24th Annual Meeting of the Association for Computational Linguistics*, 1986.

[11] C. Gussenhoven, J. t'Hart, A. Cohen, and R. Collier, "A perceptual study of intonation. an experimental-phonetic approach to speech melody," *Language*, vol. 68, p. 610, 1992.

[12] C. De Looze, I. Yanushevskaya, A. Murphy, E. O'Connor, and C. Gobl, "Pitch declination and reset as a function of utterance duration in conversational speech data," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 3071–3075.

[13] M. Farrús, C. Lai, and J. D. Moore, "Paragraph-based prosodic cues for speech synthesis applications," in *Proceedings of the International Conference on Speech Prosody*, 2016, pp. 1143–1147.

[14] L. Chen, M. J. Gales, N. Braunschweiler, M. Akamine, and K. Knill, "Integrated expression prediction and speech synthesis from text," *IEEE Journal on Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 323–335, 2014.

[15] M. Domínguez, M. Farrús, A. Burga, and L. Wanner, "The information structure—prosody language interface revisited," in *Proceedings of the 7th International Conference on Speech Prosody (SP2014), Dublin, Ireland*, 2014, pp. 539–543.

[16] M. Domínguez, M. Farrús, A. Burga, and L. Wanner, "Using hierarchical information structure for prosody prediction in content-to-speech application," in *Proceedings of the 8th International Conference on Speech Prosody (SP 2016), Boston, MA*, 2016.

[17] J. Kleinhans, M. Farrús, A. Gravano, J. M. Pérez, C. Lai, and L. Wanner, "Using prosody to classify discourse relations," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association INTERSPEECH, Stockholm, Sweden*, 2017, pp. 778–81.

[18] J. Y. Zhang, A. R. Toth, K. Collins-thompson, and A. W. Black, "Prominence prediction for super-sentential prosodic modeling based on a new database," in *Proceedings of the 5th ISCA Speech Synthesis Workshop*, 2004.

[19] K. Prahallad, A. R. Toth, and A. W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases." in *Proceedings of the INTERSPEECH*. ISCA, 2007, pp. 2901–2904.

[20] M. Schröder and J. Trouvain, "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching," *International Journal of Speech Technology*, vol. 6, pp. 365–377, 2003.

[21] L. Wanner, E. Andre, J. Blat, S. Dasiopoulou, M. Farrús, T. Fraga, E. Kamateri, F. Lingenfelser, G. Llorach, O. Martnez, G. Meditskos, S. Mille, W. Minker, L. Pragst, D. Schiller, A. Stam, L. Stellingwerff, F. Sukno, B. Vieru, and S. Vrochidis, "KRISTINA: A knowledge-based virtual conversation agent," in *Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems (PAAMS), Porto, Portugal*, 2017, pp. 284–295.