A Methodology for Generated Text Annotation for High Quality Speech Synthesis

Dimitris Spiliotopoulos and Costas Vassilakis
Department of Informatics and Telecommunications
University of the Peloponnese
Tripoli, Greece
dspiliot@uop.gr , costas@uop.gr

Dionisis Margaris
Department of Informatics and Telecommunications
University of Athens
Athens, Greece
margaris@di.uoa.gr

Kostantinos Kotis
Department of Cultural Technology and Communication,
University of the Aegean,
Lesvos, Greece
kotis@aegean.gr

***Abstract*—Natural Language Generators may generate texts that are linguistically enriched. These may result in significantly improved synthetic speech. At the same time, the generators produce pieces of plain text that may span between a single word to a full sentence. Additionally, traditional natural language generators have limited domain coverage, resulting in restricted language analysis of the generated texts. For those cases the enriched input to the speech synthesizer, required for high quality speech synthesis, can be provided by analysing the plain text. This work reports on the method for automatic domain dependent annotation of plain text, through the utilisation of the linguistic information from rich generated text. The synthetic speech from the resulting prosody models is evaluated by human participants showing annotation results for plain text quite on par with the rich generated text. This leads to improved perceived naturalness of the synthesized speech.**

Keywords—Text-to-speech, Prosody enrichment, Natural language generation, Natural language processing, Semantic feature annotation.

I. Introduction

Document-to-speech (DtS) systems, as well as other similar systems that either include a generation module or utilize generated text to synthesize speech, require rich information about the text to achieve naturalness in speech. The procedure through which the linguistic information is identified, added and associated with the text is the text annotation. Linguistically annotated text is used for many natural language processing tasks such as speech synthesis and prosody modelling. The text is annotated with specific type of information that can be derived from several linguistic analysis levels such as grammar, syntax, morphology, semantics, pragmatics, phonetics, emotion, as well as any other type of description that could prove useful.

To assign prosody, part-of-speech (POS) analysis is traditionally performed be speech synthesizers that construct the syntactic trees of the sentences [1,2]. Most general-purpose Text-to-Speech (TtS) systems involve several specific language processing systems for a list of processes, such as sentence segmentation, entity identification, and POS tagging, for the written text input. The result is the transformation of the original plain text to a rich, synthesis-aware form prior to synthesis. Due to the nature of the language processing, such analysis can suffer from high statistical error that may be due to either the inherent design and implementation of the respective language modules or language ambiguity.

To achieve naturalness, the TtS system output aims to generate high-quality speech. When the domain of the text is known (domain-dependent systems) high-quality speech is achieved due to the fact that the analysis modules are trained or otherwise designed for the specific thematic domains. However, the quality drops considerably, usually below acceptable levels, for texts of unknown domains since the analysis systems perform with lesser accuracy. Most TtS systems are modular to allow external modules or systems for analysis, however they are not designed for deep linguistic analysis, such as semantics or pragmatics, that can be used to aid synthetic speech quality. Concept-to-Speech (CtS) systems, on the other hand, include natural language generation components that produce already processed, rich, annotated text that can be used as input for the speech synthesis [3]. The generated text is error-free and annotated syntactically exhibiting full disambiguation. In addition, detailed linguistic information related to prosody may be generated that can provide considerable depth to guide synthesis. As a result, CtS systems can utilize the linguistic features from the natural language generation phase in order to produce significantly improved synthesized speech [4].

One of the major drawbacks of CtS systems is that the Natural Language Generators (NLG) are designed to operate in specific thematic domains, and thus restricted to limited domain text generation. Another real-life problem that can make the analysis more challenging is that the NLG may not always generate appropriate text output at all, due to gaps in the embedded grammar or syntax models. Another typical behaviour is that large chunks of unprocessed plain text are also produced by the generator. However, those are not processed enriched texts, but rather phrases of plain texts that are too complex to be generated and have been designed to be included in the output based on grammar generation rules. Those canned texts span from single words, to groups of words, to phrases or whole sentences that contain linguistic content, usually domain-specific, as exemplified in Fig. 1. An example of that is the MPIRO corpus [5] in which canned text counts for more than 40% of the generated text descriptions of the domain of museum exhibit descriptions. The generated speech for such texts is of variable naturalness, high for the rich generated texts and low for the canned/plain text.

978-1-7281-4959-2/19/$31.00 ©2019 IEEE

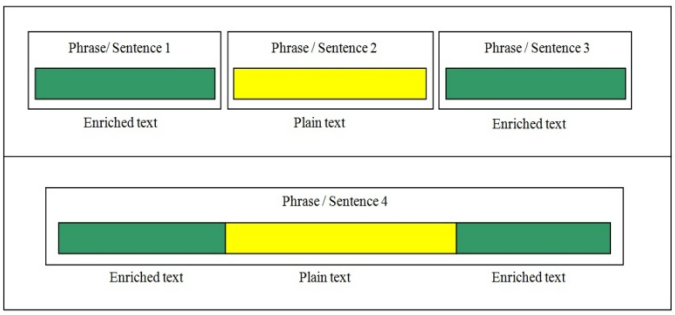

Fig. 1. Illustrated indicative sequences of phrases of (enriched and plain) generated text. Phrases or sentences of the different types of text may be intermixed (top). There are also several instances of sentences that have both types of texts (bottom)

Previous works have examined texts generated by NLGs and reported that speech synthesis quality and naturalness may be improved significantly with linguistically enriched annotated text input [6,7,8,9,10]. This is because tone generation and placements and prosodic phrasing derived from high level linguistic input result in better speech prosody than ones derived from plain text analysis [11]. This means that the standard TtS systems language processing modules may be overridden when enriched text input is present.

The main challenge that this work addresses is the uniform linguistic automatic annotation of plain generated texts using the linguistic data from the enriched generated ones. The hypothesis is that the enriched text linguistic data would provide very accurate information that could be used to train the models for the plain text automatic annotation. The advantages that would be utilised are:

1. all texts are from the same domain; therefore the meaning of certain entities remains similar to both types of text,
2. the same process of annotation (automatic and manual) would be applied to both types of texts, resulting in uniform level of analysis and
3. the plain text annotation would benefit from all the enrichment data from the generated text plus the results of the standard analysis modules.

This work explores how the levels and types of annotation of the generated enriched text may affect the annotation accuracy on the plain text. Additionally, the annotation challenges of manual verification and editing for use in speech synthesis are examined.

The aim was to achieve production of enriched text descriptions that are uniform and similar or equivalent to the ones generated by the natural language component of a CtS system. The source was both plain and/or annotated text. Both types of text may be produced from either an NLG or from a plain text document. This task necessitated the exploitation of the standard language analysis tools that are included in traditional TtS and language engineering methods to extract linguistic information from the rich generated text and use it to annotate the plain text. We report on the linguistic feature set and other types of information that is important to high quality speech synthesis and the description of the workflow and key actions that are necessary for the automatic annotation and the prosodic annotation by experts that follows. The resulting synthetic speech was evaluated by human listeners for English and Greek generated texts. The results of the differences between the unprocessed and the processed texts to the synthesized text are presented as calculated from the subjective prosody evaluation experiments.

The rest of the paper is structured as follows: section 2 overviews the background and related work, while section 3 presents the proposed approach. Section 4 presents the corpus and the experiment setup. Section 5 describes the prosodic annotation procedure based on the annotated text and section 6 presents the results of the user evaluation on the generated speech. Finally, section 7 presents the discussion of the results of the work.

II. Background and Related Work

Text generation produces texts of all sizes and types. Plain text, being one of the most common types, is also one of the hardest types of generated text to accurately feed to speech synthesizers for high quality speech output. Several works address high quality speech synthesis, in terms of signal processing [12], acoustic/prosodic features [13], segmentation [14] and syntax [15], among others [16]. The lack of such rich linguistic information on the text results in stereotypical prosody [17]. Linguistically enriched text may be utilized by the speech synthesis to achieve higher quality prosody. Ideally, a TtS synthesis system would prefer an enriched text input that would include as much information about the described text as possible. The synthesis module can then utilise such information for successful prosody construction. In the case of plain text input, the traditional speech synthesiser would employ a text analysis in order to identify and extract necessary information such as sentence and phrase breaks, part-of-speech, syntactic tree. The lexical and syntactic features are important elements for building prosody models. The sophistication and accuracy of the language analysis module greatly affect the resulting prosodic structure formulation.

The generated text by a natural language generator almost always consists of text stubs that are uneven in terms of the qualitative and quantitative amounts of linguistic information that is retained/generated. Depending on the actual design and domain feed of the generator, three types of text output can be identified: plain text, enriched-only text, and mixed text (enriched+plain). Plain text contains no extra linguistic information whatsoever, while enriched-only text usually contains lexical and syntactic information. Theoretically, this information is uniformly present and should not display any inconsistencies. In this case, the presence of error-free information renders any shallow language analysis unnecessary by the synthesiser modules. Moreover, depending on the specific generation component, semantic information, such as rhetorical relations between facts (e.g. contrast), fact-specific semantic attributes (new, partially-known, old information), or importance (also noted as “explicit emphasis”) may be exported. This is high-level linguistic information that is utilised for accurate prediction of focus prominence during speech synthesis. Semantic meta-information is analysed according to prosody-aimed features that can span one or more tokens and are used – along with lexical and syntactic features – as *prime* and *validation* factors for focus prominence assignment. Prime factors provide the state necessary for approval or disapproval of focus prominence while the presence of the

validation factors either infers the validity or differentiates the degree of focus prominence for lexical items.

Having interconnected parts of plain and enriched text hinders the resulting prosody model. The reason is twofold. The two sets of text can be very different in terms of the sentence/phrase type and size. The canned text is plain (no annotations) and usually contains larger and more complex phrase structures. The automatic language analysis that has to be applied onto the plain text to acquire the necessary lexical and syntactic description is not error-free. Apart from that, the analysis cannot involve the same lexical categories and compatible annotation with the enriched portions of text since it comes from different models. Moreover, the rhetorical relations (where available) between sentences may be compromised since the appropriate meta-information will be available only on the enriched segment of the text, therefore rendering the result of the prime and validation factor analysis untrustworthy for the larger part of the text. This happens because the content selector of the generator may include pieces of canned text in between sentences, as well as inside a single sentence containing enriched text, as depicted in Fig. 1.

III. The Proposed Approach

The proposed approach aims to tackle the following challenges:

a. Lexical and syntactic analysis of the plain text should be described by the same set of categories as the ones existing in the enriched text,

b. The annotation schemas should be compatible throughout the corpus,

c. The approach should mitigate the fact that the domain-specific content may not be covered by the standard POS analysis of the typical TtS system.

Theoretically, POS tagging is an easy processing step with adequate accuracy. For TtS prosody modelling, we are especially interested in POS since it is not only a validation factor for newness of information but also an intonation descriptor. Moreover, the POS accuracy for proper nouns and other named-entities that are actively used content words especially important for newness-based accent tone assignment [7]. The effect is very profound in limited domains where there may be several domain-specific words that are very useful to have properly annotated to aid prosody. Going beyond POS, the canned plain text disrupts the continuity of semantic relations between phrases and sentences. In plain text situations where there is no "semantic feature repair step", the successful identification of such tokens can be used to model where the missing semantics fit into the plain text phrases.

The proposed approach is as follows:

1. Extract POS information from the enriched text and optionally check and edit the annotations,
   1.1 Retrain the statistical POS tagger, using the extracted POS information to create a more accurate model for the domain. This is a crucial step for narrow domains such as the one we investigated
   1.2 Augment the lexicons, using the extracted information to add new values and append new values for existing entries. For this step, it is crucial that expert linguists append the existing entries to re-rank the multiple values according to the findings from the enriched text, since the plain text is designed to complement it both syntactically and in context.
2. If the domain is similar to the statistical POS tagger trained domain or the enriched text information coverage per text description is more than 80% of the total text (i.e. less than 20% being plain text), then use the statistical POS tagger for the plain text, otherwise use the augmented lexicons for the POS tagging. Depending on accuracy, as checked by human expert annotators, keep either as first choice. If the lexicon is kept, then the statistical should fill in for words not in the lexicons (more than one lexicon may be applicable depending on the domain). Optionally, the lexicon may be used first to create initial values and the statistical right after,
3. Update the prosody model to map the categories derived from the enriched text for better accuracy in prosody modelling.

There are several actions that can be taken to enhance the accuracy of the annotations. One is to select the domain specific words (the ones that are not covered by the lexicon) for training the statistical tagger for the plain text. Semantics, such as rhetorical relations that may be present can be corrected if breached and added in plain text. Table I shows the categories that are enriched by the proposed approach and the level of enrichment. The * denotes the situations that traditionally require human intervention or significant manual annotation.

IV. Corpus Experiments

The methodology was tested using a corpus that was generated for two languages, English and Greek, by the Ilex generator [3]. The text is generated in the form of an XML description. The corpus is a direct description of the type of generated text that was referred to in the previous paragraph.

The generated corpora contained about 53.5% plain text which was distributed in about 39% of the sentences. This meant that the plain text sentences contained more than half of the words more than the enriched text sentences. This is justified from the generator constraints for phrase and sentence size as well as other factors such as the domain authors' decisions on the plain (canned) text descriptions and associated rules. The two language corpora were about the same size. Tables II and III summarize the two corpora.

The generated enriched text contained the following annotations:

1. Part-of-speech (noun, verb, numerical, etc.),
2. Proper Noun types (names of persons, organizations, locations, artefacts, etc. were identified as "person", "organization", "location", etc.),
3. Temporal expressions,
4. Numerical quantities,

There was also partial annotation on important prosody semantic elements such as newness (whether a meaning is new to the reader or listener, applied mostly on proper nouns and adjectives), contrast, and explicitly stated emphasis.

Certain pre-processing had to be applied in order for the corpus to get annotated with prosodic information. Pre-processing mainly includes sentence type identification, named entity recognition and POS tagging. For both English and Greek texts, these processes are fully supported by existing approaches and modules: sentence and word identification are performed by a rule-based component (tokenizer) that exhibits accuracy that approaches 100% for both languages. For POS tagging, a machine learning based approach has been utilised. The POS tagging approach was based on Transformation-based Error-driven learning [18] and provided models for English (with accuracy that approaches 97%) as well as Greek (with an accuracy that approaches 80%) for generic domain processing. It is evident that the accuracy for Greek, being a highly inflective language, is not satisfactory. It is also expected to show a degraded accuracy when applied to different domains than the ones it was trained for.

We extracted the linguistic information from the enriched text and re-trained the statistical POS tagger. At the same time, in the enriched text, about 20% of the nouns in the English corpus and 27% of the nouns in the Greek corpus were not covered by the lexicons. They were added to both lexicons and, for the purposes of this work, were validated by a professional linguist. The POS tags were also validated by two professional linguists.

Accuracy and recall were measured for validated and non-validated instances for plain, enriched and plain+enriched generated texts over the same instances for the trained statistical and the lexicon-based taggers. Tables IV and V show the results for the two language corpora. The tables show how the full text, plain (annotated with the enriched information) and enriched based models perform on their respective types of text. It is evident that the enriched text generated annotations lead to accurate results for all selected procedures (statistical and lexicon-based).

In Tables IV and V, *Pat* stands for "POS all text", *Pglb* for "POS generated (lexicon-based)", *Pgs* for "POS generated (statistical)", *Pp* for "POS plain", *Pplb* for "POS plain (lexicon-based)", *Pps* for "POS plain (statistical)", *Pe* for "POS enriched", *Pelb* for "POS enriched (lexicon-based)" and *Pes* for POS enriched (statistical).

TABLE I. COMPLETENESS OF ANNOTATION PER GRAMMATIC LEVEL

	BEFORE PROCESSING			**AFTER PROCESSING**		
Level	**Total**	**Enriched**	**Plain**	**Total**	**Enriched**	**Plain**
Lexical	Partial	Full	None	Full	Full	Full (A)
Sentence	Partial	Full	None	Full	Full	Full (A)
Syntactic	Partial	Full*	None	Full*	Full	Full (A or M)
Semantic	Partial	Limited	None	Full*	Full*	Full*

TABLE II. GENERATED TEXT CORPUS (ENGLISH LANGUAGE)

Generated text	**Sentences No / %**	**Tokens No / %**	**Words No / %**	**Words per sentence**
Enriched	251 / 56.9	2949 / 47.0	2635 / 46.8	10.5
Plain	190 / 43.1	3329 / 53.0	3000 / 53.2	15.8
Total	441 / 100.0	6278 / 100.0	5635 / 100.0	12.8

TABLE III. GENERATED TEXT CORPUS (GREEK LANGUAGE)

Generated text	**Sentences No / %**	**Tokens No / %**	**Words No / %**	**Words per sentence**
Enriched	267 / 65.9	3012 / 47.8	2673 / 46.1	10.0
Plain	201 / 34.1	3418 / 53.2	3088 / 53.9	15.4
Total	468 / 100.0	6430 / 100.0	5731 / 100.0	12.2

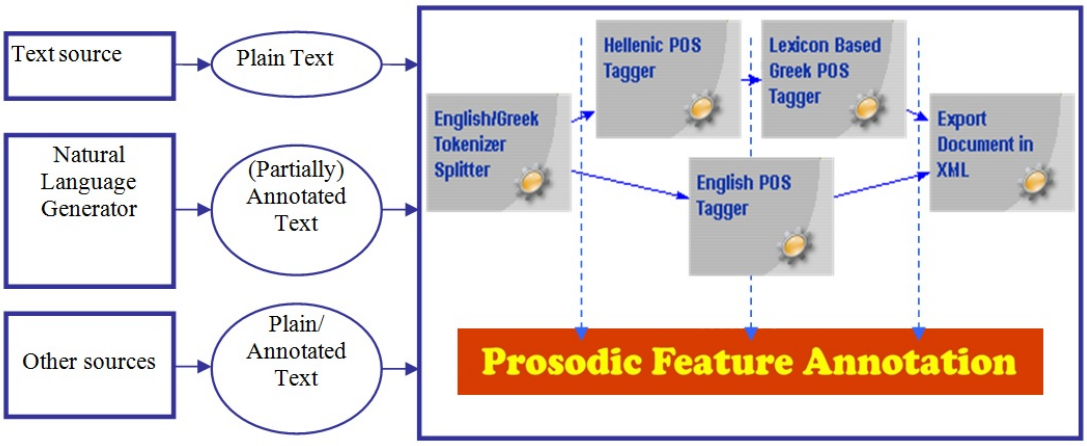

Fig. 2. Standard annotation workflow for generated text for speech synthesis

V. Prosodic Annotation

For the speech synthesis, ideally, specific information is valuable. The following are standard input for high-quality speech synthesizers:

1. Word list,
2. Syntax & POS (full syntactic trees of sentences, phrases, words),
3. Focus information (between phrases and between words),
4. Phrase boundaries, i.e. location and types of boundaries (major/minor, parentheses, etc.),
5. Common pause marks information (paragraph pause, blank line pause),
6. Special marks information (e.g. a "title" phrase requires special pause after the utterance),
7. Semantic descriptions (new/given information, contrast).

Traditional TtS systems, in general, accept plain text as input, using internal specialized algorithms for the generation of speech-related natural language information prior to synthesis. Nevertheless, the algorithms that are typically employed for such tasks are neither adequately powerful nor specialized to comprehensively identify speech-related information about the multitude of linguistic phenomena from the plain text, providing thus text analysis of limited depth; this applies also to the derived

TABLE IV. POS annotation results (English)

Annotated Sources	**Target text for automatic annotation (Precision / Recall)**								
POS	Pat	Pglb	Pgs	Pp	Pplb	Pps	Pe	Pelb	Pes
all text (validated)	.965 / .932								
all text (non-validated)		.901 / .767	.877 / .746						
plain (validated)				.945 / .918					
plain (non-validated)					.874 / .771	.849 / .749			
enriched (validated)							.988 / .988		
enriched (non-validated)								.935 / .762	.911 / .742

TABLE V. POS annotation results (Greek)

Annotated Sources	**Target text for automatic annotation (Precision / Recall)**								
POS	Pat	Pglb	Pgs	Pp	Pplb	Pps	Pe	Pelb	Pes
all text (validated)	.952 / .952								
all text (non-validated)		.910 / .761	.899 / .752						
plain (validated)				.916 / .901					
plain (non-validated)					.877 / .751	.887 / .792			
enriched (validated)							.979 / .979		
enriched (non-validated)								.936 / .752	.916 / .766

descriptions. The provision of pre-processed annotated text as input to the speech synthesizer is a valuable alternative. The major advantage that the enriched text of that kind exhibits over plain text is that it retains semantic, sentence-level structural, syntactic and discourse-level linguistic information in the form of tags in the mark-up. Each of the above categories of linguistic information is represented by sets of features, which can be exploited to improve the quality of the generated prosody in speech synthesis. The sets of features utilized may vary depending on the type as well as the domain of text, to achieve optimal performance. The efficient realization of this procedure necessitates the availability of automated analysis and annotation components for the most stages of language analysis (Fig. 2).

A breakdown of the identifiable distinct processes is:

- Word/Sentence identification and segmentation.
- Shallow syntactic analysis (part-of-speech tagging and noun-phrase identification).
- Creation/export to appropriate XML format description.
- Insertion/annotation of prosodic features.

It is possible to realize fully automated analysis for all the above processes, except for the last one. The first two were described in earlier sections. The XML export is an extension of the SOLE-ML description that caters for the prosodic description [19]. It was originally built as an annotation scheme for CtS synthesis, used as mark-up for the enriched text output of the generator. A module for automated extraction to produce the augmented XML description based on SOLE-ML has been realized and incorporated in the process flow. Linguistic phenomena such as rhetorical relations, anaphoric references, and deixis are especially difficult to automatically detect using only plain text as input. This is usually a manual task for experts resulting in flexible and broad feature sets that may be used as meta-information and also maintaining compliance to speech-oriented XML mark-up for both editing and export.

For the manual annotation of prosodic features, we used a visual editor (Fig. 3). The prosodic features that were annotated were contrast, definition, disjunction, emphasis (explicit), exemplification, newness, non-newness, similarity, yes-no-question, wh-question.

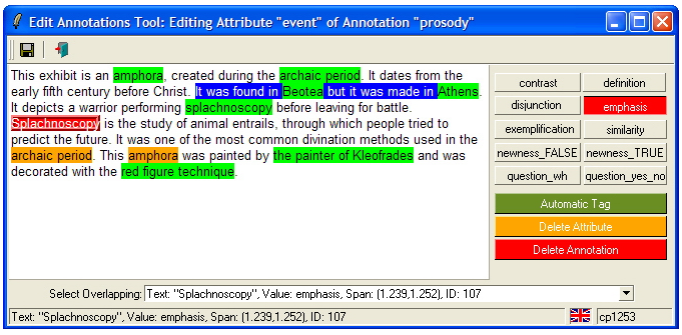

Fig. 3. Prosodic feature annotation

The prosodic feature annotation was performed by three linguists. They all informally reported that the annotation process was much easier when using fully POS annotated text. Accuracy in the token assignment was also much higher than normal (low number of manual errors). That was justified from the fact that that there may be more than one feature associated with each word or set of words (or phrase) required for successful description used for prosody modelling. Working with fully annotated text allowed the experts to select the correct overlapping annotations as well as nesting for each token or groups of tokens.

VI. User Evaluation

The results of the proposed method of annotation were evaluated at the utilisation point, speech prosody, by 10 participants (age *M*=23.4, *SD*= 4.2) that were asked to listen to random sections of synthesized unprocessed and processed (annotated) plain and enriched text. The users were all native Greek speakers, while two of them were also native English speakers. They were asked to listen to random synthetic speech queues generated from plain and enriched text and report their feedback based on the scales from ITU-T Rec. P.85 that refer to prosodic evaluation [20]. These were *Overall impression* (MOS), *Voice pleasantness* (PLT), *Accentuation* (ACCT), *Listening effort* (LSTE), *Comprehension problems* (COPR), and *Acceptance* (ACCP). The random speech queues presented to the listeners 25 of each selection from processed and unprocessed plain and enriched, to a total of 100 per language. The participants recorded their feedback on the Likert 1-5 scale. The two languages were evaluated by the participants during separate sessions with a long break in between.

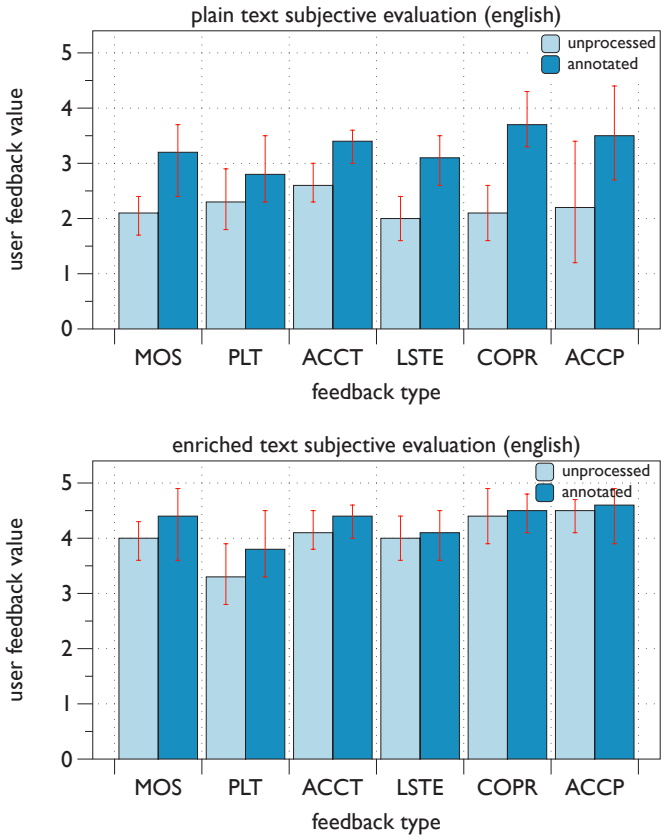

Fig. 4. User evaluation for English synthetic speech

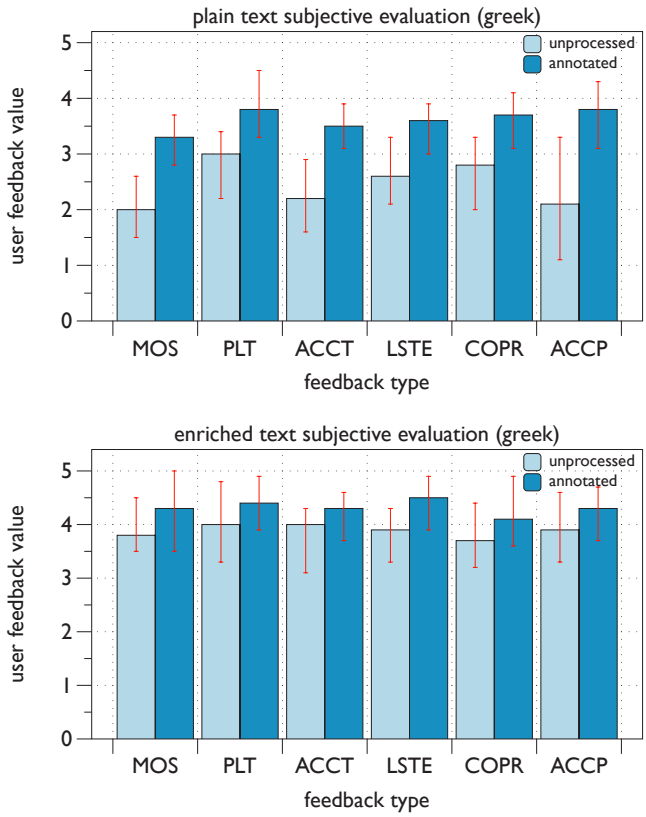

Fig. 5. User evaluation for Greek synthetic speech

Figures 4 and 5 present the user evaluation. The bars depict the average feedback scores along with the standard deviation. The findings were similar for both languages. Although, the speech synthesis can handle plain text, the comparative evaluation from the listeners resulted in a low overall impression and acceptance. In fact, for all feedback types, the unprocessed plain text was rendered 20-30% lower than the same processed text, for both languages. The annotation of the plain texts resulted in significantly higher values, much nearer to the enriched text before processing.

The prosody from the enriched text was quite accurate to begin with. With the additional annotations, the naturalness was ranked even higher, scoring an additional 6-10% on average for all feedback types.

VII. Discussion

For high quality speech synthesis, intonational focus point placement and type cannot always be inferred by POS and phrase type linguistic information. This happens because those are not only influenced by syntax but also by semantics and pragmatic aspects. For prosody modelling in speech synthesis, these aspects can be utilised for computation, deduction and verification of focus prominence and the text corpus is appropriately enriched to take them into account. An ideal input for a speech synthesiser would be an annotated text with focus information. Since this is very hard for a natural language generator to export, the next best thing would be error-free lexical and syntactic information and consistent rhetorical information.

The challenge is that generators produce as much plain canned text as they do enriched text phrases. The plain text sentences are larger and more complex. The generator either cannot generate those kinds of sentences automatically or requires a lot of authoring for rarely-reusable domain dependent lexicon words. Due to the nature and the limits of the text generators (namely grammar and domain of application), big data processing approaches [21] are not suitable for linguistic annotation.

The proposed methodology extracts information from the generated enriched text and uses it to automatically annotate the plain text. This enables the prosody annotation, whether manual or automatic, to utilise uniform enriched text for annotation of prosodic information. This leads to higher annotation accuracy on all aspects that are necessary for the training of prosody models for high naturalness in speech synthesis.

Future work includes the use of the proposed methodology on the speech synthesis for accessible interfaces [22-24], enhancing usability [25,26] and analysis of social media text [27-31]. Finally, we are planning to include collaborative filtering techniques in order to achieve automatic preferences from inter-annotator agreement scores [32-42].

References

[1] P. Taylor, A. Black, and R. Caley, “The architecture of the festival speech synthesis system,” Proceedings of the 3rd ESCA Workshop on Speech Synthesis, pp.147–151, 1998.

[2] T. Dutoit, “An Introduction to Text-to-Speech Synthesis,” Kluwer Academic Publishers, Dordrecht, 1997.

[3] M. O’Donnel, C. Mellish, J. Oberlander and A. Knott, “ILEX: An architecture for a dynamic hypertext generation system,” Natural Language Engineering, vol.7(3), pp. 225-250, 2001.

[4] J. Hitzeman, A. Black, P. Taylor, C. Mellish and J. Oberlander, "On the Use of Automatically Generated Discourse-Level Information in a Concept-to-Speech Synthesis System," Proceedings of the 5th International Conference on Spoken Language Generation, pp. 2763-2768, 1998.

[5] I. Androutsopoulos, D. Spiliotopoulos, K. Stamatakis, A. Dimitromanolaki, V. Karkaletsis, and C. Spyropoulos, “Symbolic Authoring for Multilingual Natural Language Generation,” Proceedings of the Hellenic Conference on Artificial Intelligence, pp. 131-142, 2002.

[6] S. Pan, K. McKeown and J. Hirschberg, “Exploring features from natural language generation for prosody modeling,” Computer Speech and Language, vol. 16, pp. 457-490, 2002.

[7] G. Xydas, D. Spiliotopoulos and G. Kouroupetroglou, "Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora," IEICE Transactions of Information and Systems, vol. E88-D(3), pp. 510-518, 2005.

[8] G. Xydas, D. Spiliotopoulos and D. Kouroupetroglou, Modeling Prosodic Structures in Linguistically Enriched Environments. Proceedings of the 7th International Conference on Text, Speech and Dialogue, pp. 521-528, 2004.

[9] G. Xydas, D. Spiliotopoulos and G. Kouroupetroglou, “Modelling Emphatic Events from Non-Speech Aware Documents in Speech Based User Interfaces”. Proceedings of the 10th International Conference on Human-Computer Interaction, pp. 806-810, 2003.

[10] D. Spiliotopoulos, G. Xydas and G. Kouroupetroglou, “Diction Based Prosody Modeling in Table-to-Speech Synthesis”. Proceedings of the 8th International Conference on Text, Speech and Dialogue, pp. 294-301, 2005.

[11] A. Black and P. Taylor, “Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input,”

Proceedings of the 3rd International Conference on Spoken Language Processing, pp.715–718, 1994.

[12] L. Violante, P. Rodrigue and A. Gravano, "Improving speech synthesis quality by reducing pitch peaks in the source recordings," Proceedings of NAACL-HLT 2013, pp. 502-506, 2013.

[13] R. H. Galvez, S. Benus, A. Gravano, and M. Trnka, "Prosodic facilitation and interference while judging on the veracity of synthesized statements," Proceedings of Interspeech 2017, pp. 2331–2335, 2017.

[14] W. Yang and K. Georgila, "Automatic Detection of Unnatural Word-Level Segments in Unit-Selection Speech Synthesis," Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition Understanding, pp. 289-294, 2011.

[15] N. Obin, P. Lanchantin, M. Avanzi, A. Lacheret-Dujour and X. Rodet, "Toward improved HMM-based speech synthesis using high-level syntactical features," Proceedings of the 2010 Speech Prosody, pp. 2000-2004, 2010.

[16] M. Schroeder, "Expressive Speech Synthesis: Past, Present, and Possible Futures," Affective Information Processing, pp. 111 –116, 2009.

[17] S. Padda, N. Bhalla and R. Kaur, "A Step towards Making an Effective Text to Speech Conversion System," International Journal of Engineering Research and Applications, vol. 2(2), pp. 1242 –1244, 2012.

[18] E. Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging," Computational Linguistics, vol. 21, pp. 543-565, 1995.

[19] J. Hitzeman, A. Black, C. Mellish, J. Oberlander, M. Poesio and P. Taylor, "An annotation scheme for Concept-to-Speech synthesis", Proceedings of the 7th European Workshop on Natural Language Generation, pp. 59-66, 1999.

[20] ITU-T Rec. P.85, "A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices," International Telecommunication, 1994.

[21] H. Zen, A. Senior and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," Proceedings of 38th International Conference on Acoustics, Speech, and Signal Processing, pp. 1-5, 2013.

[22] D. Spiliotopoulos, G. Xydas, G. Kouroupetroglou, V. Argyropoulos and K. Ikospentaki, "Auditory Universal Accessibility of Data Tables using Naturally Derived Prosody Specification," Universal Access in the Information Society, vol. 9, pp. 169-183, 2010.

[23] D. Spiliotopoulos, P. Stavropoulou and G. Kouroupetroglou, "Acoustic Rendering of Data Tables using Earcons and Prosody for Document Accessibility," Proceedings of the 5th International Conference on Universal Access in Human-Computer Interaction, pp. 587-596, 2009.

[24] A. Pino, G. Kouroupetroglou, H. Kacorri, A. Sarantidou and D. Spiliotopoulos, "An open source / freeware assistive technology software inventory", Proceedings of the 12th International Conference on Computers Helping People with Special Needs, pp. 178–185, 2010.

[25] D. Spiliotopoulos, P. Stavropoulou and G. Kouroupetroglou, "Spoken Dialogue Interfaces: Integrating Usability," Proceedings of the Symposium of the Austrian HCI and Usability Engineering Group, pp. 484-499, 2009

[26] D. Spiliotopoulos, E. Tzoannos, P. Stavropoulou, G. Kouroupetroglou and A. Pino, "Designing user interfaces for social media driven digital preservation and information retrieval," Proceedings of the 13th International Conference on Computers Helping People with Special Needs, pp. 581-584, 2012.

[27] D. Antonakaki, D. Spiliotopoulos, C.V. Samaras, S. Ioannidis and P. Fragopoulou, "Investigating the Complete Corpus of Referendum and Elections Tweets," Proceedings of the IEEE/ACM Conference on Advances in Social Networks Analysis and Mining, pp. 100-105, 2016.

[28] G. Schefbeck, D. Spiliotopoulos and T. Risse, "The Recent Challenge in Web Archiving: Archiving the Social Web," Proceedings of the International Council on Archives Congress, pp. 20-24, 2012.

[29] D. Antonakaki, D. Spiliotopoulos, C.V. Samaras, P. Pratikakis, S. Ioannidis and P. Fragopoulou, "Social media analysis during political turbulence," PloS one, vol. 12(10), pp. 1-23, 2017.

[30] T. Risse, E. Demidova, S. Dietze, W. Peters, N. Papailiou, K. Doka, Y. Stavrakas, V. Plachouras, P. Senellart, F. Carpentier, A. Mantrach, B. Cautis, P. Siehndel and D. Spiliotopoulos, "The ARCOMEM Architecture for Social and Semantic-driven Web Archiving", Future Internet, vol. 6(4), pp. 688–716., 2014

[31] E. Demidova, N. Barbieri, S. Dietze, A. Funk, H. Holzmann, D. Maynard, N. Papailiou, W. Peters, T. Risse and D. Spiliotopoulos, "Analysing and Enriching Focused Semantic Web Archives for Parliament Applications", Future Internet, vol. 6(3), pp. 433-456, 2014

[32] D. Margaris and C. Vassilakis, "Exploiting Internet of Things Information to Enhance Venues' Recommendation Accuracy," Service Oriented Computing & Applications, vol. 11(4), pp. 393-409, 2017.

[33] D. Margaris and C. Vassilakis, "Enhancing User Rating Database Consistency through Pruning," Transactions on Large-Scale Data and Knowledge-Centered Systems, vol. XXXIV, pp. 33–64, 2017.

[34] D. Margaris and C. Vassilakis, "Improving Collaborative Filtering's Rating Prediction Quality by Considering Shifts in Rating Practices," Proceedings of the 19th IEEE International Conference on Business Informatics, pp. 158-166, 2017.

[35] D. Margaris, C. Vassilakis and P. Georgiadis, "Knowledge-Based Leisure Time Recommendations in Social Networks," Current Trends on Knowledge-Based Systems: Theory and Applications, pp. 23-48, 2017.

[36] D. Margaris, C. Vassilakis and P. Georgiadis, "Adapting WS-BPEL scenario execution using collaborative filtering techniques," Proceedings of the 7th IEEE International Conference on Research Challenges in Information Science, pp. 174-184, 2013.

[37] D. Margaris and C. Vassilakis, "Exploiting Rating Abstention Intervals for Addressing Concept Drift in Social Network Recommender Systems," Informatics, vol. 5(2), Article no. 21, 2018.

[38] D. Margaris, C. Vassilakis and P. Georgiadis, "Recommendation information diffusion in social networks considering user influence and semantics," Social Network Analysis and Mining, vol. 6(1), 108, pp. 1-22, 2016.

[39] D. Margaris, C. Vassilakis and P. Georgiadis, "Query personalization using social network information and collaborative filtering techniques," Future Generation Computer Systems, vol. 78(1), pp. 440-450, 2018.

[40] D. Margaris and C. Vassilakis, "Improving Collaborative Filtering's Rating Prediction Quality in Dense Datasets, by Pruning Old Ratings," Proceedings of the 22nd IEEE Symposium on Computers and Communications, pp. 1168-1174, 2017.

[41] D. Margaris and C. Vassilakis, "Improving Collaborative Filtering's Rating Prediction Accuracy by Considering Users' Rating Variability," Proceedings of the 4th IEEE International Conference on Big Data Intelligence and Computing, pp. 1022-1027, 2018.

[42] D. Margaris and C. Vassilakis, "Enhancing Rating Prediction Quality through Improving the Accuracy of Detection of Shifts in Rating Practices," Transactions on Large-Scale Data- and Knowledge-Centered Systems, vol. XXXVII, pp. 151-19, 2018.