The SP2 SCOPES Project on Speech Prosody

György Szaszák[1,2], Tamás Gábor Csapó[2], Philip N. Garner[1], Branislav Gerazov[3], Zoran Ivanovski[3], Géza Németh[2], Bálint Tóth[2], Milan Sečujski[4], Vlado Delić[4]

***Abstract*—This is an overview of a Joint Research Project within the Scientific co-operation between Eastern Europe and Switzerland (SCOPES) Program of the Swiss National Science Foundation (SNFS) and Swiss Agency for Development and Cooperation (SDC). Within the SP2 SCOPES Project on Speech Prosody, in the course of the following two years, the four partners aim to collaborate on the subject of speech prosody and advance the extraction, processing, modeling and transfer of prosody for a large portfolio of European languages: French, German, Italian, English, Hungarian, Serbian, Croatian, Bosnian, Montenegrin, and Macedonian. Through the intertwined four research plans, synergies are foreseen to emerge that will build a foundation for submitting strong joint proposals for EU funding.**

***Keywords*—Speech prosody, Speech recognition, Text-to-speech synthesis**

I. Introduction

A. Significance of prosody

PROSODY, comprising intonation, dynamics and rhythm, is one of the most significant building blocks of spoken language and serves to carry information about the discourse function, salience, and speaker attitude and emotion, [1], [2], [3]. The lack of proper prosody can make the speech sound unnatural and hard to follow, even though it might be fully intelligible. Thus prosody generation modules have a crucial role in text-to-speech (TTS) synthesis systems since the beginning, [4]. This is especially true for current state-of-the-art TTS systems, whose focus is set not only on sounding natural, but also on showing expressions [5], [6], [7], [8]. Thus prosody generation is a part of TTS systems that is under intense research, [9], [10], [8]. With the emergence of speech-to-speech translation (S2ST) technology, treatment — extraction, event detection and implied modelling — of prosody has become necessary on the automatic speech recognition (ASR) side as well [11]. Another application of prosody on the ASR side is the syntactic analysis based on spoken language, which can be exploited in disambiguation and analysis before the translation process [12].

B. Prosody extraction

The three basic acoustic features of prosody are F0 (pitch), energy and duration, and they can sometimes be complemented with other, less frequently used features such as jitter, shimmer, HNR [13], etc. In the realisation of prosodic constituents, like stress or intonation, the three basic features may interact and have different importance. The contribution of different features is somewhat language dependent [14] (for example, duration is an important cue of stress in American English, whereas in Hungarian, F0 is believed to be the dominant cue of stress with duration playing almost no role in it). Extracting energy is a basic task, which can usually be carried out without complications. Extraction of duration patterns, on the other hand, may pose problems, especially if the underlying segmental structure (phone segmentation) of the speech signal is unknown. The biggest challenge remaining in this field is the accurate extraction of F0 [15]. Although several somewhat reliable algorithms are known, the F0 estimate is typically corrupted by doubling/halving errors [16]. The F0 contour is not continuous, but human perception of it is capable of keeping track as if it were continuous. Recent research has found that using a continuous F0 contour can be advantageous [17]. In speech technology applications, the F0 contour is often interpolated to overcome problems caused by discontinuity. An alternative to this approach has recently been proposed based on probabilistic features for F0 extraction [18].

C. Prosody models

In order to produce prosody for speech synthesis, or to analyze it on the ASR side, it is necessary to have a model that is able to capture and generate the characteristics of prosody. Many attempts of building models of pitch have been carried out, and they can be put in two main categories: the ones that directly model pitch, and the ones trying to simulate the pitch production process. If the former category counts numerous models [19], [20], [21], [22], the latter has only a few models [23], [24]. Most of the models can be used to analyze and represent pitch, however they cannot be used directly with existing pitch generation models such as those used in the HMM synthesis framework. Another approach to pitch modeling is the target approximation by [25]. A review of the main approaches and models is presented in [9]. Only several of these models try to take into account the different levels of prosody and there is no agreement on the suitable number of levels to be used. Finally, most of the existing models use only low level (i.e. short term) variations to describe prosody.

D. Prosody prediction

There has been significant research in the proper modeling of stress patterns and the prediction of prosody from text in the field of text-to-speech synthesis [26], [27], [28]. These methods have the goal of automatically assigning stress patterns to any given input text in order to output synthesized speech with human-like prosody. As in everyday human-human communication, proper prosody is of great importance. The position of stressed words can depend on a number of circumstances (e.g. context, mood, emphasis of the sentence), and predicting natural prosody

The authors would like to thank the Swiss National Science Foundation (SNFS) and Swiss Agency for Development and Cooperation (SDC) for their financial support of the SP2: SCOPES Project on Speech Prosody, within the Scientific co-operation between Eastern Europe and Switzerland (SCOPES) Program, No. IZ73Z0_152495.

[1]Idiap Research Institute, Martigny, Switzerland

[2]Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics (BME-TMIT), Budapest, Hungary

[3]Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University of Skopje (FEEIT), Skopje, Macedonia

[4]Faculty of Technical Sciences, University of Novi Sad (FTS-UNS), Novi Sad, Serbia

patterns is one of the most relevant stages in speech synthesis. Previous solutions applied pre-defined rules [29] or used data-driven intonation prediction methods [26], [28].

E. Salient event detection and crosslingual prosody transfer

Prosody transfer during speech-to-speech translation (S2ST) is a recent research area with increasing importance. So far, there have been a few approaches to consider source speech prosody, especially salient events in S2ST. A salient event is regarded to be a prominence or focus which does not result from syntactic constraints, but is rather a product of effects at the semantic or pragmatic level. Whereas canonical prosody (e.g. a stress pattern, intonation or timing patterns) can be generated based on text (morphology and syntax), this is not the case for salient events, which occur specifically and relatively unpredictably according to the speaker's intention. [11] used a bilingual speech corpus to perform unsupervised clustering to find intonation clusters of the source speech in order to map them to corresponding intonation clusters in the target speech. However, this approach presumes that some prosodic isomorphism exists. To avoid this presumption, [30] integrated the generation of accent information into statistical translation models using factored translation models (see [31]). Both approaches restrict to intonation and do not consider either rhythm or loudness. Furthermore they intend to detect each and every pitch accent and translate it to the output speech. The relatively rare salient prosodic events are not considered explicitly or treated in a special way.

The detection of salient prosodic events is related to prosodic labeling, which has been a research topic for many years [32]. The aim of this labeling is to segment (by means of ASR techniques) and annotate speech signals that can be used to train statistical prosody models for speech synthesis or syntactic analysis [12].

Various classifiers for automatic prosody labeling of read speech have been proposed. Most often, ToBI-annotated (or with alternative annotation for tones and break indices, similar to the ToBI philosophy as in [33]) corpora are used where labels are often mapped to broad categories to circumvent sparsity problems. [34] presented an intonation and stress classifier and alignment system using prosodic unit models instead of discrete prosodic event models. [35] report that using solely acoustic features for a prosodic classifier performed roughly as well as using a combination of acoustic, lexical and syntactic features. The same holds for conversational speech [36] where acoustic features alone led to good performance.

In the closely related SIWIS project [37], it is argued that considering only salient prosodic events for prosody transfer is advantageous for several reasons:

- Salient events can usually be detected reliably because they are prominent. If an event is not detected because it is not salient enough (or it is detected but is considered to be not prominent enough) it is regarded not requiring special treatment.
- For the transfer of an event bound to some word(s) in source language (L1), we have to find the corresponding word(s) in the target language (L2). In case we use a rule-based translation system this is easy to achieve, because mappings are known. When using a data-driven translation system, the corresponding position in the L2 text can be found from the (generally ambiguous) translations of the individual words.
- Salient prosodic events are not only expressed by means of F0 movement (as in other works), but also by intensity and duration (lengthening and pausing). All these properties of such an event should be considered.

The same paradigm is to be followed in the present project too, in order to obtain the most complex prosodic coverage possible when transferring prosody by still maintaining a flexible, not overly complicated system architecture (cf. [31]).

II. Partners and Collaboration

This research is a joint research of 4 partners, funded by the SCOPES mechanism of the Swiss National Science Foundation. The partners are:

- Idiap Research Institute (Idiap), Martigny, Switzerland
- Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics (BME-TMIT), Budapest, Hungary
- Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University – Skopje (FEEIT), Skopje, Macedonia
- Faculty of Technical Sciences, University of Novi Sad (FTS-UNS), Novi Sad, Serbia

The form of collabration is a joint project, which means that each of the 4 partners is working on more or less individual tasks, whereby common points are identified and other partners interact to reach a higher level of integration. This scheme also promotes technological transfer and the filling of the technological gap between European countries, a priority of the SCOPES funding mechanism.

III. Research Directions

The general aim of the joint project in terms of scientific research is to advance prosody extraction, processing and modeling by involving a large portfolio of European languages (English, French, German, Hungarian, Italian, Serbian, Macedonian) where beside Western languages, Central- and Eastern European languages are also represented. Detection of salient prosodic events and investigation of crosslingual transfer is also of interest. Salient events are especially frequent in conversational / spontaneous speech, the latter being an active research topic nowadays. This aim is similar to that of the SIWIS project, however, within the present context, experiments are extended to several new languages, like Slavic Serbian and Macedonian (Indo-European family) and Finno-Ugrian Hungarian (the most spoken European language of non Indo-European origin). The goal is to advance conversational modeling, prosody modification techniques and crosslingual transfer of prosody. Another aspect assessed is prosody — and especially stress – prediction based on text. Although numerous approaches exist for this purpose, there is a lack of methods that would assign stress patterns to textual input obtained from human speech in a prosody transplantation way. For this task, it is needed to create new databases, as well as integrate available methods and advanced models of the stress patterns of human speech.

In the following, some research directions are presented briefly. Most of the individual tasks involve a close collaboration and scientific exchange among the partners, especially if techniques involving several languages (transfer,

comparison, universal modeling) are developed. Individual efforts are also expected to converge and contribute to reach the common research goals.

A. Idiap

Idiap is already working on prosody via the projects SIWIS and RECOD. These projects aim to build prosodic models for the Swiss languages and for coding respectively, and also to develop prosodic event detection and prosody transfer. The technical objective of the SP2 collaboration is twofold. One is to identify whether these models generalise to other languages, notably those of the project partners. The other is to identify which benefits the expertise of the other partners can bring in the context of the existing projects, in addition to extending the language portfolio with Central and South-East European languages.

a) Improvements in prosodic event detection: The goal is to develop methods for automatic detection of both canonical (e.g. syntactic stress) and salient prosodic events (e.g. the positions of contrastive or emphatic stress, major hesitations or pauses, locally reduced speaking rate, etc.) in the source signal. The detection results can be used to assign tags or labels to the ASR output word sequence helping either translation or deep linguistic analysis based on spoken utterances. The planned research regarding prosody extraction includes the evaluation of the recently released tool [18] that gives a probabilistic, continuous F0 estimate. The prosodic classifier and aligner presented in [34] and reported to have robust pause and stress detection capabilities is to be tested when using probabilistic F0 and also ported from Hungarian to Swiss languages (to French at least, as a fixed stress might be a bounding condition). Partners already have the necessary speech data to perform these experiments, some extension with proper prosodic labels in a semi-supervised manner is foreseen.

b) Investigation of prosodic models: We are interested in whether common prosodic models (such as the Fujisaki model [29]) are actually universal models or, at least, to what extent are they universal or if there is a possibility to universalise them to a certain extent. This task involves the active collaboration of all partners.

As discussed earlier, every language has its own prosodic characteristics, and in the context of crosslingual prosody adaptation it appears as a difficulty to overcome, because the idea of using a joint prosody model is excluded. However, similarities can be found across languages, namely when the languages are close such as English and German. For instance (in these languages), it is commonly admitted that a declarative sentence tends to have an overall decreasing pitch, while interrogation is marked by a raising pitch. An analysis of these is planned in order to evaluate to what extent universalisation is feasible in prosody modelling.

c) Crosslingual prosody transfer: Crosslingual transfer of prosody is to be assessed for languages involved (selected from the official languages in Partner countries). This considerably extends the work done in SIWIS for new languaguges of the Indo-European language family and Finno-Ugrian Hungarian from a different language family. A special focus is placed on salient event transfer, allowing for richer semantic-pragmatic analysis or transfer when used in S2ST applications.

Translation involves changed word order, and often even the number of words is different between the source and target language. In current project, translation is treated as a black box application, however, supposing a rule-based operation. In this case, word mappings are known and event transfer is feasible using the appropriate transfer models. Especially salient event transfer will be analysed for language pairs involved, complementing to the SIWIS work, lead also by Idiap.

d) Coordination: Idiap is also responsible for the coordination of the entire project.

B. BME-TMIT

The main objective of BME-TMIT within the SP2 project is to investigate similarities and differences in prosody extraction, modeling and prediction between Hungarian (being a Finno-Ugrian language) and other European languages (Indo-European languages). The Hungarian partner is particularly interested in experiments with stress and the F0 component of prosody.

e) Database extensions and other resources: The Hungarian partner BME-TMIT is engaged in a prosodically labelled textual database extension: the rule-based stress and prosody prediction method of the Profivox system [38] will be used as a basis. It will be used for assigning stress patterns for 2000 Hungarian sentences, and the rule-based stress prediction will be corrected manually by listening to synthesized sentences. The result will be a high precision textual database with correct stress markings. Two types of stress will be differentiated: stressed word and unstressed word. Beside this planned database, the speech recordings of the same 2000 sentences are available with 10 Hungarian speakers [39], which will be used for experiments of prosody extraction.

f) Comparison of rule-based and data-driven stress prediction: The Hungarian partner will carry on a comparison of rule-based and data-driven stress prediction. The database with the stress markings can be used to train data-driven methods for the prediction of stress patterns and assign them to textual input. In order to investigate the effectiveness of data-driven stress prediction it will be compared to previous rule-based methods (e.g. a deterministic, superpositional prosody model for Hungarian implemented in the Profivox system [40]).

g) Analysis and re-synthesis of speech with a source-filter based vocoder: With pitch modification techniques and vocoder methods (e.g. [41]) it is possible to change the prosody and therefore the stress patterns of a sentence by analyzing the speech signal, modifying the fundamental frequency (F0) and resynthesizing the sentence. For this task, a reliable F0 detection method is required, therefore it is planned to use the recent F0 estimation tool of Idiap [18]. To investigate the accuracy and robustness of this research tool, the Hungarian partner will test it on large amount of Hungarian data including frequent creaky voice. The vocoder can be applied in HMM-based speech synthesis as well. Adaptation of this vocoder system to salient event generation is foreseen.

h) Investigation of the language dependency of stress prediction: As the stress patterns of sentences is highly language dependent (e.g. in Hungarian stress is always on the first syllable of a words, whereas in other languages this is not true), it is usually necessary to create different methods for different languages and compare the language dependency of stress prediction.

The rule-based and data-driven methods developed for Hungarian stress prediction might be suitable for other European languages as well. It is a question, however,

whether the methods are suitable for crosslingual prosody transfer.

C. FEEIT

The primary focus of research of FEEIT within the scope of this project is the development of a high quality prosody model for spoken Macedonian. Research into this area has already been started [42], [43], [44], [45], and has led to the development of an automatic intonation generation algorithm for Macedonian [46], [47]. In order to develop a high-quality prosody model a dedicated corpus is to be created and annotated, to which the algorithms for prosody extraction developed within the scope of this project are to be applied. The jointly developed prosody model and prosody prediction algorithms are then to be evaluated on the Macedonian prosody database. Finally crosslingual prosody transfer is to be investigated in line with the other languages represented in the project. The Macedonian partner will not only offer its insights in the specifics of the Macedonian language, but will also actively participate in the development of algorithms and methods used throughout the project.

i) Improvements in prosodic event detection: The Macedonian team will take an active part in the development and improvement of efficient digital signal processing algorithms for pitch and rhythm extraction to be used in the process of prosody analysis of the annotated corpora. The usability of these algorithms for the various languages covered within the scope of the project will be evaluated.

j) Investigation of prosodic models: There is a clear necessity for the development of a more sophisticated model for Macedonian prosody. The models developed so far for Macedonian prosody offer only a simple description of the underlining parameters of its patterns. More importantly, they have been developed only on the scale of intonation phrases, thus excluding the crucially important prosodic fluctuations at the level of words and syllables. It is the goal of FEEIT in accordance with Idiap, to work on developing a unifying prosody model that can be adapted to the various languages the partners in this project represent, but even further on a pan-European and global scale.

k) Database creation and analysis: To facilitate testing of the improved prosody models in their proficiency to model prosody patterns in spoken Macedonian, a corpus will be recorded and annotated for analysis. In the process of creation of the Macedonian prosody corpus, key know-how will be used from BME-TMIT. The algorithms developed in the scope of this project are to be used to extract the prosody from the audio recordings, by performing amplitude, pitch and phone duration extraction. "Blind" segmentation algorithms developed by FEEIT [48] and FTS-UNS [49], will be used in conjunction with the small-vocabulary ASR system developed for Macedonian [50], to automate the segmentation of the corpus.

l) Adaptation of algorithms for prosody prediction: The developed Macedonian TTS System includes simple rule-based prosody prediction algorithm which needs to be further developed to improve the quality of the synthesized speech output. To this end, the knowledge gained by BME-TMIT and adapted to the particularities of the Macedonian language, will be used to improve prosody generation under the unifying model developed with Idiap.

m)Crosslingual prosody transfer: This direction is part of the joint efforts to which FEEIT will contribute with testing prosody transfer possibilities for Macedonian.

D. FTS-UNS

Having developed a prosody model for Serbo-Croatian and integrated it into a speech synthesizer for two standard varieties of the language [51], the immediate objective of the FTS-UNS partner is to examine the possibilities of extending the model to both the remaining varieties of Serbo-Croatian and to other more or less kindred European languages, including the languages of the project partners. The general objective of the FTS-UNS partner is to establish a network of partnerships aimed at the exchange of knowledge and good practices, joint development of general-purpose tools and extension of the language portfolio.

n) Prosody modeling at the sentence level: The synthesized speech obtained by the current version of the AlfaNum TTS, which is by far the best available TTS for any of the standard varieties of Serbo-Croatian, is highly intelligible and reasonably natural. However, as the prosodic annotation scheme it uses is restricted to lexical accent types (or, equivalently, lexical tonal patterns) and specification of prosodic events such as minor or major phrase breaks and salient prosodic events, the intonation of the entire sentence may appear rather "flat" to a listener, although individual intonation phrases are pronounced with highly satisfactory intonation [51]. Namely, one of the directions of future research will be concerned with the comprehensive analysis of sentence prosody, related to its large-scale constituents, adding expressiveness to the synthesized speech. The research will be based on the available speech corpora and new corpora which are to be recorded and annotated, including a new expressive TTS corpus that is currently being annotated [52]. The research will be oriented on the development of a more sophisticated prosody model along the lines of [53].

o) Prosody extraction and prediction using neural networks: The algorithms for prosody extraction and prediction, currently based on classification and regression trees, will be enhanced by the introduction of recurrent language networks. More sophisticated methods for the prediction of salient prosodic events will also be investigated using both techniques. The research will include the automatic identification of salient prosodic events (including contrastive or emphatic stress, pauses and significant local changes in speech), training of machine learning systems to learn relevant relations between salient prosodic events and features such as part-of-speech, position in relation to other words and phrase breaks, the number of times that the word has occurred in recent history etc. The trained systems will then be used for the assignment of salient prosodic events to new utterances.

p) Extension of the established approach to other kindred languages: Another direction of future research will include the extension of the established approach to Bosnian and Montenegrin, which will provide valuable new information on the issues of language dependency and crosslingual prosody transfer. Namely, although Serbian, Croatian, Bosnian and Montenegrin (incipient) are mutually intelligible standard varieties of a single pluricentric language (Serbo-Croatian), they nevertheless exhibit significant differences and require individual treatment when it comes to speech synthesis. Within this project a new annotated speech corpus will be developed for Bosnian,

along the lines of the existing corpora for Serbian and Croatian (annotation of lexical accent types i.e. tonal patterns, as well as specification of prosodic events such as minor or major phrase breaks and prosodic prominence) [54].

q) Development of general-purpose software tools for speech prosody treatment: Finally, the team from the Faculty of Technical Sciences, University of Novi Sad, will also participate in the development of software tools for speech prosody treatment. In this development the team will build upon the resources and algorithms it has developed so far, including general-purpose signal processing library [55], pitch-extraction algorithm [56], and a blind speech segmentation tool based on features extracted by the existing ASR modules [49], which was successfully applied to Hebrew, a language significantly different from Serbo-Croatian [57]. The team will also rely on a number of the existing speech and language resources for Serbo-Croatian [54].

IV. Conclusion

The project presented in this paper is focused on speech prosody the extraction and synthesis of pitch, duration and intensity of speech. The primary goal of the project is to advance conversational modeling, prosody modification techniques and the crosslingual transfer of prosody. Prosody is an important, yet not well understood aspect of speech, limiting its inclusion in automatic speech recognition (ASR) and text to speech synthesis (TTS) systems. To amend this a model of speech is required that will go beyond surface acoustics, describing real natural cues and production mechanisms. In a speech to speech translation system, such cues must come from the utterance being translated, extracted using ASR. These cues must then be properly translated to the target language and forwarded to the TTS system.

The capital contribution of this joint project upon its completion will be the advancement in prosody extraction, processing, modeling and transfer for a large portfolio of European languages: French, German, Italian, English, Hungarian, Serbian, Croatian, Bosnian, Montenegrin, and Macedonian. The project integrates four research plans and identifies synergies that will allow the transition from independent work to a homogeneous collaboration. Each of the partners has a set of research directions that intertwine with those of the other partners. It is foreseen that through the course of achieving the set individual and joint goals of the project, a state will be reached in which the partners will be able to submit a strong joint proposal for EU funding (including FP8 and Horizon 2020).

References

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2 ed.* Prentice Hall, 2008.

[2] N. Campbell, "Developments in corpus-based speech synthesis: Approaching natural conversational speech," in *IEICETIS*, vol. E88-D, 3, March 2005.

[3] D. O'Shaughnessy, "Modern methods of speech synthesis," *IEEE Circuits and Systems Magazine*, vol. 7, no. 3, pp. 6–23, 2007.

[4] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers, April 1997.

[5] E. Székely, T. G. Csapó, B. Tóth, P. Mihajlik, and J. Carson-berndsen, "Synthesizing expressive speech from amateur audiobook recordings," in *SLT*, Miami, Florida, USA, 2012, pp. 297–302.

[6] M. Tatham and K. Morton, *Developments in Speech Synthesis*. John Wiley & Sons Ltd., 2005.

[7] J. Pitrelli, R. Bakis, E. Eide, R. Fernandez, W. Hamza, and M. Picheny, "The ibm expressive text-to-speech synthesis system for american english," in *TSAP*, vol. 14 (4), July 2006, pp. 1301–1312.

[8] M. Bulut and S. Narayanan, "Expressive speech synthesis using a concatenative synthesizer," in *ICSLP 2002*, Denver, Colorado, USA, September 2002.

[9] P. Taylor, *Text-to-speech synthesis*. Cambridge University Press, 2009.

[10] J. Adamek, "Neural networks controlling prosody of Czech language," Magister Thesis, Univerzita Karlova v Praze, Matematicko-fyzikaln fakulta, 2002.

[11] P. Aguero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *Proc. of ICASSP 2006*, vol. 1, 2006, pp. 700–705.

[12] G. Szaszák and A. Beke, "Exploiting Prosody for Automatic Syntactic Phrase Boundary Detection in Speech," *Journal of Language Modeling*, vol. 1, pp. 143–172, 2012.

[13] P. Boersma, "Accurate short-term analysis if the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences, Amsterdam*. University of Amsterdam, 1993, no. 17, pp. 97–110.

[14] D. Hirst and A. Di Cristo, *Intonation systems: a survey of twenty languages*. Cambridge University Press, 1998.

[15] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 365–375, 2009.

[16] K. Murray, "A study of automatic pitch tracker doubling/halving "errors"," in *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, vol. 16. Association for Computational Linguistics, 2001, pp. 1–4.

[17] K. Yu and S. Young, "Continuous F0 modelling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2011.

[18] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102–105, January 2013.

[19] G. Bailly and B. Holm, "SFC: A trainable prosodic model," *Speech Communication*, vol. 46, no. 3, pp. 348–364, 2005.

[20] J. Pierrehumbert, "Synthesizing intonation," *JASA*, vol. 70, pp. 985–995, 1981.

[21] D. Hirst and R. Espesser, "Automatic modelling of fundamental frequency using a quadratic spline function." *Travaux de l'Institut Phontique d'Aix*, pp. 75–85, 1993.

[22] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[23] G. Kochanski, C. Shih, and H. Jing, "Quantitative measurement of prosodic strength in mandarin," *Speech Communication*, vol. 41, no. 4, pp. 625–645, 2003.

[24] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing. acoustical analysis and physiological interpretations," *Dept. for Speech, Music and Hearing, Tech. Rep*, 1981.

[25] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in mandarin and english as a process of target approximation," *JASA*, vol. 125, pp. 405–424, January 2009.

[26] A. Raux and A. W. Black, "A unit selection approach to f0 modeling and its application to emphasis," in *Proc. ASRU 2003*, 2003, pp. 700–705.

[27] J. van Santen, A. Kain, E. Klabbers, and T. Mishra, "Synthesis of prosody using multi-level unit sequences," *Speech Communication*, vol. 46, no. 3-4, pp. 365–375, 2005.

[28] F. C. Díaz, J. van Santen, and E. R. Banga, "Integrating phrasing and intonation modelling using syntactic and morphosyntactic information," *Speech Communication*, vol. 51, no. 5, pp. 452–465, 2009.

[29] H. Fujisaki, "Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing," in *The Production of Speech*, P. Mac Neilage, Ed. Springer, New York, 1983, pp. 39–55.

[30] R. Sridhar, S. Bangalore, and S. Narayanan, "Factored translation models for enriching spoken language translation with prosody," in *Proceedings of Interspeech*, 2008, pp. 2723–2726.

[31] P. Koehn and H. Hoang, "Factored translation models," in *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, p. 868876.

[32] A. Rosenberg, "Automatic Detection and Classification of Prosodic Events," Ph.D. dissertation, Columbia University, USA, 2009.

[33] F. Gallwitz, H. Niemann, E. Nth, and W. Warnke, "Integrated recognition of words and prosodic phrase boundaries," *Speech Communication*, vol. 36, pp. 81–95, 2002.

[34] K. Vicsi and G. Szaszák, "Using prosody to improve automatic speech recognition," *Speech Communication*, vol. 52, no. 5, pp. 413–426, 2010.

[35] J. Jeon and Y. Liu, "Syllable-level prominence detection with acoustic evidence," in *Proceedings of Interspeech*, 2010.

[36] R. Sridhar, A. Nenkova, S. Narayanan, and D. Jurafsky, "Detecting prominence in conversational speech: pitch accent, givenness and focus," in *Proc. Speech Prosody*, 2008, pp. 380–388.

[37] P. N. Garner, H. Bourlard, B. Pfister, E. Wehrli, R. Clark, and J. Yamagishi, "SIWIS: Spoken Interaction with Interpretation in Switzerland." [Online]. Available: http://www.idiap.ch/project/siwis/

[38] G. Olaszy, G. Németh, P. Olaszi, and G. Kiss, "Profivox - A Hungarian text-to-speech system for telecommunications applications," *International Journal of Speech Technology*, vol. 3, no. 3-4, pp. 201–215, 2000.

[39] G. Olaszy, "Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai [Development and services of a Hungarian precisely labeled and segmented, parallel speech database] (in Hungarian)," *Beszédkutatás 2013 [Speech Research 2013]*, pp. 261–270, 2013.

[40] G. Olaszy, G. Németh, and P. Olaszi, "Automatic prosody generation-a model for hungarian." in *Proc. Eurospeech*, 2001, pp. 525–528.

[41] T. G. Csapó and G. Németh, "A novel codebook-based excitation model for use in speech synthesis," in *IEEE CogInfoCom*. Kosice, Slovakia: IEEE, Dec. 2012, pp. 661–665.

[42] B. Gerazov and Z. Ivanovski, "Analysis of intonation in the Macedonian language for the purpose of text-to-speech synthesis," in *EAA EUROREGIO 2010*, Ljubljana, Slovenia, September 2010.

[43] ——, "Analysis of intonation dynamics in Macedonian for the purpose of text to speech synthesis," in *TELFOR 2010*, Belgrade, Serbia, November 2010.

[44] ——, "Analysis of extracted pitch contours across speakers for intonation modelling in TTS synthesis," in *5th International Symposium on Communications, Control, and Signal Processing ISCCSP 2012*, Rome, Italy, May 2012.

[45] B. Gerazov, Z. Ivanovski, and R. Bilibajkic, "Modeling Macedonian intonation for text-to-speech synthesis," in *DOGS 2010*, Iriski Venac, Serbia, December 2010.

[46] B. Gerazov and Z. Ivanovski, "Generation of pitch curves for Macedonian text-to-speech synthesis," in *6th Forum Acusticum*, Aalborg, Denmark, June/July 2011.

[47] ——, "Prosody generation module for Macedonian text-to-speech synthesis," in *AES 130th Convention*, London, UK, May 2011.

[48] B. Gerazov, M. Bogdanov, and Z. Ivanovski, "Segmentation of speech based on the undecimated wavelet transformation (in Macedonian)," in *ETAI 2009*, Ohrid, Macedonia, September 2009.

[49] V. Delić, M. Sečujski, N. Jakovljević, M. Janev, R. Obradović, and D. Pekar, *Speech Technologies for Serbian and Kindred South Slavic Languages*. SCIYO, 2010, ch. 9. Advances in Speech Recognition, pp. 141–164.

[50] B. Gerazov and Z. Ivanovski, "Prototype automatic speech recognition system for a voice dialing application for Macedonian (in Macedonian)," in *Summer Symposium on Electronics and Signal Processing LEOS 2012*, Mavrovo, Macedonia, September 2012.

[51] M. Sečujski, D. Pekar, and N. Jakovljević, "Automatic prosody generation for Serbo-Croatian speech synthesis based on regression trees," in *INTERSPEECH 2011, Florence, Italy*, 2011, pp. 3157–3160.

[52] M. Sečujski, S. Ostrogonac, S. Suzić, and D. Pekar, "Speech database production and tagset design aimed at expressive text-to-speech in Serbian," in *DOGS 2014, Novi Sad, Serbia*, 2014.

[53] S. Godjevac, *Transcribing Serbo-Croatian Intonation*. Oxford Linguistics, 2005, pp. 146–171.

[54] V. Delić, M. Sečujski, N. Jakovljević, D. Pekar, D. Mišković, B. Popović, S. Ostrogonac, M. Bojanić, and D. Knežević, "Speech and language resources within speech recognition and synthesis systems for Serbian and kindred south Slavic languages," in *SPECOM 2013, Pilsen, Czech Republic*, ser. LNCS, 2013, vol. 8113, pp. 319–326.

[55] D. Pekar and R. Obradović, "C++ signal processing library - slib 1.0 (in Serbian)," in *DOGS 2000, Sremski Karlovci, Serbia*, 2000, pp. 67–70.

[56] D. Knežević, D. Pekar, I. Milošević, and N. V. Sedlar, "An adaptive system for the detection of fundamental frequency based on AMDF (in Serbian)," in *DOGS 2008, Kelebija, Serbia*, 2008, pp. 32–34.

[57] N. Jakovljević, D. Mišković, D. Pekar, M. Sečujski, and V. Delić, "Automatic phonetic segmentation for a speech corpus of Hebrew," in *11. INFOTEH-JAHORINA, Jahornina, Bosnia & Herzegovina*, 2012.