Design of a Speech Corpus for Research on Cross-Lingual Prosody Transfer

Milan Sečujski[1], Branislav Gerazov[2], Tamás Gábor Csapó[3], Vlado Delić[1], Philip N. Garner[4], Aleksandar Gjoreski[2], David Guennec[5], Zoran Ivanovski[2], Aleksandar Melov[2], Géza Németh[3], Ana Stojković[2], and György Szaszák[3]

[1] Faculty of Technical Sciences, University of Novi Sad, Serbia
secujski@uns.ac.rs
[2] Faculty of Electrical Engineering and Information Technologies, University of Ss. Cyril and Methodius, Skopje, Macedonia
gerazov@feit.ukim.edu.mk
[3] Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
[4] Idiap Research Institute, Martigny, Switzerland
[5] IRISA Research Institute, Rennes, France

Abstract. Since the prosody of a spoken utterance carries information about its discourse function, salience, and speaker attitude, prosody models and prosody generation modules have played a crucial part in text-to-speech (TTS) synthesis systems from the beginning, especially those set not only on sounding natural, but also on showing emotion or particular speaker intention. Prosody transfer within speech-to-speech translation is a recent research area with increasing importance, with one of its most important research topics being the detection and treatment of salient events, i.e. instances of prominence or focus which do not result from syntactic constraints, but are rather products of semantic or pragmatic level effects. This paper presents the design and the guidelines for the creation of a multilingual speech corpus containing prosodically rich sentences, ultimately aimed at training statistical prosody models for multilingual prosody transfer in the context of expressive speech synthesis.

Keywords: prosody, speech corpus, speech synthesis, speech-to-speech translation

1 Introduction

The ambition of current state-of the art systems is not only to produce intelligible and natural sounding speech, but also to approach humans in their ability to convey emotion or a particular speaker intention [1,2,3,4]. For that reason, prosody modeling and prediction are arguably the most important research challenges in the domain of text-to-speech (TTS) synthesis [4,5,6]. The relevance of prosody for automatic speech recognition (ASR) has also begun to gain appreciation, particularly with the advent of speech-to-speech (STS) translation systems

[7]. Just as humans disambiguate spoken utterances and give them a proper linguistic interpretation relying on prosody, automatic systems now attempt do the same, which can be of particular importance in the context of STS [8]. Furthermore, by taking sentence intonation and other prosodic features into account, salient prosodic events, which represent intentional speaker deviations from the canonical prosody, can be detected and, if properly modelled, can be carried over to the target language and introduced into synthesized speech, with the ultimate goal of preserving the original speaker intention. However, the treatment of salient prosodic events is a complex task, since their realization constitutes an interplay between the basic prosody features (intonation, timing and dynamics), just as is the case with canonical prosody, which is generally determined by the morphology and syntax of the utterance (e.g. by stress patterns and ordering of sentence constituents).

Since prosody transfer within speech-to-speech translation is a recent research area, there have so far been relatively few approaches to analyse source speech prosody in terms of salient events and carry them over to the target language. The assumption that there exists some isomorphism between the source and the target language greatly simplifies the problem. For instance, the research in [7], using a bilingual speech corpus as training material, was based on performing unsupervised clustering of intonation patterns in the source speech in order to directly map them to corresponding intonation clusters in the target speech. However, a general case where such an assumption cannot be made requires a more high-level approach. In [9] the generation of pitch accent information was integrated into statistical translation models using factored translation models [10], in order to avoid possibly erroneous reconstruction of prosody of the target utterance based on the translated text only. However, besides focusing on the intonation contour and excluding other prosodic features from consideration, both approaches are based on the detection of each and every pitch accent and translating them to the output speech, rather than explicitly considering salient prosodic events which occur relatively unfrequently.

The modeling and treatment of salient prosodic events is closely related to prosodic labeling, i.e. annotating speech corpora for prosodic events (stress, accent, boundary between prosodic constituents, emphasis etc.). Prosodically annotated corpora are an indispensable tool for training statistical prosody models for a range of applications including speech synthesis or syntactic analysis of spoken utterances [8]. However, the construction of such corpora is an extremely time-consuming task, requiring a lot of manual effort, which makes such corpora relatively scarce and prompts the need for the development of automatic prosodic labeling techniques [11]. To this date, a number of various classifiers for automatic prosodic labeling of speech have been proposed (cf. e.g. [12,13,14]), based on annotation systems such as Tone and Break Indices (ToBI) [15] or other conventions for marking tones and breaks (cf. e.g. [16]), but their accuracy is still below the one that can be achieved by expert humans. This paper presents the design and the guidelines for the creation of a multilingual speech corpus containing prosodically rich sentences, representing an invaluable resource for

the research in the domain of cross-lingual prosody transfer in the context of expressive speech synthesis. The corpus has been created within the research project "SP2: SCOPES Project on Speech Prosody" supported by the Swiss National Science Foundation [17], covers 5 languages at the moment, and to the best knowledge of the authors, represents the only existing multilingual corpus specifically aimed at supporting the research into salient prosodic events and their cross-lingual transfer.

The remainder of the paper is organized as follows. Section 2 will present the content of the speech corpus in more detail and discuss the motives behind several choices that have been made. Section 3 will present the annotation guidelines and present several characteristic examples. Section 4 will briefly illustrate the utility of the corpus with an example research based on it, and Section 5 will conclude the paper with an outline of the future work.

2 Contents of the SP2 Speech Corpus

At the moment, the SP2 Speech Corpus contains sections covering English, French, Hungarian, Serbian and Macedonian, and each section contains recordings from one or two speakers so far, amounting to 7 speakers in total.[6] Following the existing guidelines for new contributions, the corpus can be easily extended to new speakers and new languages.

The set of sentences for a single speaker contains 50 prosodically rich sentences, with the same text translated into different languages. Each utterance has one or more words marked in bold to indicate emphasis. When translating the text into a new language, care was taken to preserve the original meaning of the sentence, but just as importantly, to preserve the emphasis in the translation without signaling it by other means such as a particular choice of words. For instance, for the Serbian sentence:

Đorđe$_{[n.\ \text{George}]}$ im$_{[pron.\ \text{to them}]}$ je$_{[aux.v.]}$ to$_{[pron.\ \text{about it}]}$ saopštio$_{[v.\ \text{told}]}$.

the translation into English "**George** told them about it" would be preferable to a translation that introduces a cleft sentence, such as "It was **George** who told them about it." The sentences are divided into the following 5 groups of 10 sentences:

- Emphasis on a single word. ("It turned out that it was a **fake** gun".)
- Emphasis early in the sentence. ("**Money** is what I like the most".) This specific case is treated separately in order to give a better insight into post-focus compression [18,19], i.e. the perceptible reduction of pitch range and intensity after prosodic focus.
- Emphasis marking an explicit contrast. ("Since he cannot **buy** it, he's going to **rent** it".) This section is expected to provide an insight into the differences in prosodic realization of the opposed syntactic constituents in various languages.

[6] The SP2 Speech Corpus can be downloaded from `http://gitdipteam.feit.ukim.edu.mk/gerazov/sp2_database_specom2016`, and contributions are welcome.

- Emphasis marking an explicit contrast in a question. ("Are you **emotional** or **rational**?")
- Emphasis as a result of semantic focus on a relatively large constituent. ("**It was because she felt so lonely** that she decided to move".) This section is expected to provide an insight into the speaker dependence of focus projection, i.e. the degree of variability with which different speakers map the semantic focus on a certain constituent into emphasis or pitch accent on particular words [20].

Each speaker was required to deliver:

- the described 50 sentences *with* particular emphasis on the words or parts of sentences marked in bold,
- the same 50 sentences *without* particular emphasis on the marked word or parts of sentences, to the degree to which it is reasonably possible, having in mind that in some sections, especially ones dealing with explicit contrast, it can be difficult to pronounce a particular sentence without emphasis, as emphasis "comes naturally".

The set of recordings for each speaker thus contains 100 utterances. The number of the sentences per speaker is arguably too small for the corpus to be directly used for training statistical prosody models, but it offers a possibility to study inter-speaker variability in using prosodic cues to signal emphasis in a particular language, as well as the relations between their use in different languages in a number of typical situations.

3 Annotation Guidelines

The existing speaker sets have been annotated with *Praat* [21], using the following interval tiers:

- **Emphasis**. The only mandatory tier, in which the emphasized word(s) are marked with '+', while other words are not marked. If the word is pronounced with an unusually strong emphasis, '++' is used instead. Clearly, not all words marked in bold in the text get a '+' or '++', but only ones actually emphasized. For each word marked with '+' or '++' in the emphasized utterance, there is a corresponding '(+)' or '(++)' in the neutral utterance (the non-emphasized counterpart), indicating the position of the corresponding word.
- **Contrast**. This is a semantic tier, which marks the opposing sentence constituents in sentences with explicit contrast (e.g. "Instead of getting a **rest**, I got **tired**"). Here, words in contrast ("rest" and "tired") are marked with '1' and '2' respectively. In some cases, where more than one element is emphasized on either of the opposing sides, multiple '1' or '2' tags are assigned. If a word is marked with a '1' or '2' in the emphasized utterance, it carries the same tag in the neutral utterance, regardless of the fact that it is may not be actually emphasized there.

- **Words**. This tier indicates boundaries between words, which are given in their orthographic forms in order to be matched with the text more easily.
- **Syllables**. This tier indicates boundaries between syllables, which are also given in their orthographic forms.
- **Lexical stress**. This tier marks lexically stressed syllables with a '+'. If the speaker stressed a different syllable than the one required by the standard pronunciation, a syllable actually stressed is marked with a '+' (in general, at least for some languages and particular words, there can be more than one acceptable location of the lexical stress).
- **Lexical tone**. This tier is applicable only to tonal languages or languages with pitch accent, and indicates the tone or pitch accent of a particular syllable (according to the conventions adopted for the langauge in question).
- **Phones**. This tier indicates phone boundaries and gives a phonetic transcription in SAMPA format. The purpose of this tier is to enable a more detailed analysis of pitch contours, since stress is usually related primarily to the vowel in the syllable.

The following point tier can also be used:

- **Breaks**. This tier indicates the positions of phrase breaks which significantly affect pitch in either of the two versions of the utterance. The purpose of this tier is to indicate possible sources of major pitch variations which are not due to emphasis. Unless otherwise specified for a particular language, such breaks are indicated by 'B' in both versions of the utterance even if their impact is significant in only one of them.

The following example (Fig. 1) shows the full annotation of the following Macedonian sentence (version with emphasis): "Сите мислеа дека тој **знаел** за заговорот." ("Everybody thought that he **knew** about the plot."). The contrastive stress is not marked, as there is none in this example. Similarly, lexical tones are not applicable to Macedonian, so this tier is empty, as well as the phrase breaks tier.

4 Example Research

In the course of the SP2 project several sections of the SP2 Speech Corpus have been used in research focused on salient prosodic event analysis and detection. Specifically we have looked at how emphasis is communicated in the three dimensions of prosody, through the comparison between emphasized and non-emphasized renditions of the same utterance. In the English language, both syllable duration [22] and energy [23] were seen as indicative of emphasis. Based on their analysis, emphasis detection algorithms were designed and evaluated using the SP2 Speech Corpus. Moreover, an adapted version of our Weighted Correlation Atom Decomposition (WCAD) based intonation modelling algorithm [24,25] was used to decompose the energy contour, achieving results in emphasis detection [26] comparable to the state-of-the-art [27]. The database is currently being used for the design of more sophisticated emphasis detection algorithms, as well as cross-lingual transfer of emphasis.

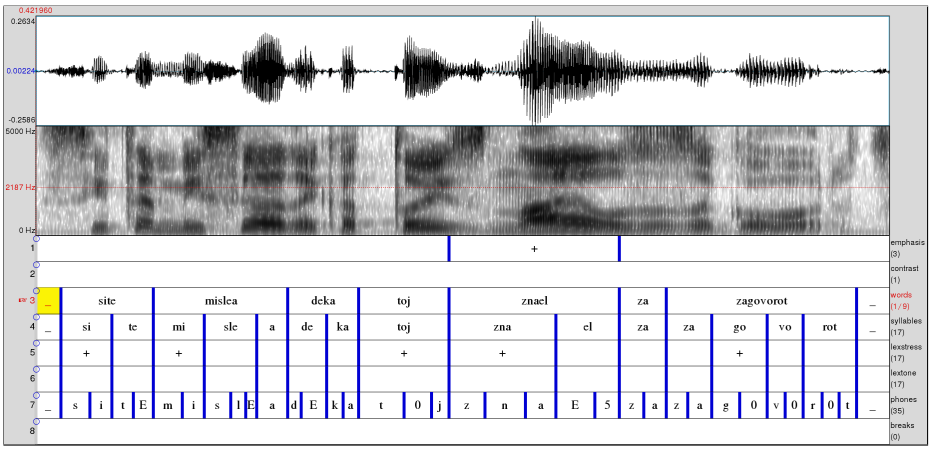

Fig. 1. Annotation of an emphasized sentence in Macedonian. In the corresponding non-emphasized sentence, the absence of emphasis would be indicated by a '(+)' marker on the emphasis tier, positioned at the corresponding word ('знаел').

5 Conclusions and Future Work

The prosodically rich SP2 Speech Corpus has been specifically designed for the research in salient prosodic event detection and their cross-lingual transfer. This is an area of research gaining particular importance with the introduction of STS translation systems which aim at conveying not only the information contained in *what* was said but also in *how* it was said. The corpus in its current form covers 5 languages and includes voices of 7 speakers, each having delivered 50 pairs of unemphasised-emphasised utterances, divided into 5 categories based on the type and/or location of emphasis. Our team has, thus far, used the corpus to successfully design and evaluate emphasis detection algorithms. It is our intention that the corpus should be of use for research conducted by the whole scientific community. Moreover, owing to well-defined guidelines for preparing contributions to the corpus, it is our hope that the community will help the corpus to expand to other languages soon.

Acknowledgments

The authors would like to acknowledge the support of the Swiss National Science Foundation via the research project "SP2: SCOPES Project on Speech Prosody".

References

1. Székely, E., Csapó, T.G., Tóth, B., Mihajlik, P., Carson-Berndsen, J.: Synthesizing expressive speech from amateur audiobook recordings. In: IEEE Workshop on Speech and Language Technology (SLT), Miami, FL, USA, pp. 297–302 (2012)

2. Tatham, M., Morton, K.: Developments in Speech Synthesis. John Wiley & Sons Ltd. (2005)
3. Pitrelli, J., Bakis, R., Eide, E., Fernandez, R., Hamza, W., Picheny, M.: The IBM expressive text-to-speech synthesis system for American English. IEEE Transactions on Audio, Speech, and Language Processing, vol. 14(4), 1301–1312 (2006)
4. Bulut, M., Narayanan, S., Syrdal, A.: Expressive speech synthesis using a concatenative synthesizer. In: 7th International Conference on Spoken Language Processing (ICSLP), Denver, CO, USA (2002)
5. Taylor, P.: Text-to-speech synthesis. Cambridge University Press (2009)
6. Adamek, J.: Neural networks controlling prosody of Czech language. Master Thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (2002)
7. Aguero, P., Adell, J., Bonafonte, A.: Prosody generation for speech-to-speech translation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 700–705 (2006)
8. Szaszák, G., and Beke, A.: Exploiting prosody for automatic syntactic phrase boundary detection in speech. Journal of Language Modeling, vol. 1, pp. 143–172 (2012)
9. Sridhar, R., Bangalore, S., Narayanan, S.: Factored translation models for enriching spoken language translation with prosody. In: INTERSPEECH, pp. 2723–2726 (2008)
10. Koehn, P., Hoang, H.: Factored translation models. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 868-–876 (2007)
11. Rosenberg, A.: Automatic detection and classification of prosodic events. Ph.D. dissertation, Columbia University, NY, USA (2009)
12. Jeon, J., Liu, Y.: Syllable-level prominence detection with acoustic evidence. In: INTERSPEECH, pp. 1772–1775 (2010)
13. Vicsi, K., Szaszák, G.: Using prosody to improve automatic speech recognition. Speech Communication, vol. 52(5), pp. 413–426 (2010)
14. Sridhar, R., Nenkova, A., Narayanan, S., Jurafsky, D.: Detecting prominence in conversational speech: pitch accent, givenness and focus. In: 4th Conference on Speech Prosody, Campinas, Brazil, pp. 380–388 (2008)
15. Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S.: The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (ed.) Prosodic Typology – The Phonology of Intonation and Phrasing, pp. 9–54. Oxford University Press, Oxford, UK (2005)
16. Gallwitz, F., Niemann, H., Nöth, E., Warnke, W.: Integrated recognition of words and prosodic phrase boundaries. Speech Communication, vol. 36, pp. 81–95 (2002)
17. Szaszák, G., Csapó, T.G., Garner, P., Gerazov, B., Ivanovski, Z., Németh, G., Tóth, B., Sečujski, M., Delić, V.: The SP2 SCOPES Project on Speech Prosody. In: Digital Speech and Image Processing (DOGS), Novi Sad, Serbia, pp. 9–14 (2014)
18. Chen, S.-W., Wang, B., Xu, Y.: Closely related languages, different ways of realizing focus. In: INTERSPEECH, pp. 1007–1010 (2009)
19. Botinis, A., Fourakis, M., Gawronska, B.: Focus identification in English, Greek and Swedish. In: 14th International Congress of Phonetic Sciences, pp. 1557–1560 (1999)
20. Selkirk, E.: Sentence prosody: intonation, stress, and phrasing. In: Goldsmith, J.: (ed.) The Handbook of Phonological Theory, pp. 550–569. Blackwell, Cambridge, MA, USA (1999)

21. Boersma, P.: *Praat*, a system for doing phonetics by computer. Glot International, vol. 5(9/10), 341–345 (2001)
22. Melov, A., Gerazov, B., Ivanovski, Z.: Towards extracting the global component from the syllable duration contour for emphatic word detection. In: 3rd International Acoustics and Audio Engineering Conference TAKTONS (2015)
23. Stojkovic, A., Gerazov, B., Ivanovski, Z.: Emphatic word detection based on relative phoneme energies within syllables. In: 12th International Conference ETAI (2015)
24. Honnet, P.-E., Gerazov, B., Garner, P.N.: Atom decomposition-based intonation modelling. In: IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP (2015)
25. Gerazov, B., Honnet, P.-E., Gjoreski, A., Garner, P.: Weighted correlation based atom decomposition intonation modelling. In: INTERSPEECH (2015)
26. Gjoreski, A., Gerazov, B., Ivanovski, Z.: Atom-decomposition based analysis for the purpose of emphatic word detection. In: 12th International Conference ETAI (2015)
27. Cernak, M., Honnet, P.-E.: An empirical model of emphatic word detection. In: INTERSPEECH (2015)