RSS-TOBI - a Prosodically Enhanced Romanian Speech Corpus

RSS-TOBI - A Prosodically Enhanced Romanian Speech Corpus

Tiberiu Boroș[1], Adriana Stan[2], Oliver Watts[3], Stefan Daniel Dumitrescu[1]

[1]Research Institute for Artificial Intelligence "Mihai Drăgănescu", Bucharest, Romania

[2]Speech Processing Group, Technical University of Cluj-Napoca, Romania

[3]Centre for Speech Technology Research, University of Edinburgh, United Kingdom

E-mail: tibi@racai.ro, adriana.stan@com.utcluj.ro, owatts@inf.ed.ac.uk, sdumitrescu@racai.ro

Abstract

This paper introduces a recent development of a Romanian Speech corpus to include prosodic annotations of the speech data in the form of ToBI labels. We describe the methodology of determining the required pitch patterns that are common for the Romanian language, annotate the speech resource, and then provide a comparison of two text-to-speech synthesis systems to establish the benefits of using this type of information to our speech resource. The result is a publicly available speech dataset which can be used to further develop speech synthesis systems or to automatically learn the prediction of ToBI labels from text in Romanian language.

Keywords: text-to-speech synthesis, Romanian language, ToBI

1. Introduction

Text-to-speech (TTS) synthesis is an important component of the Spoken Language Processing, aiming at providing human-computer interaction using speech-enabled user interfaces. And although throughout the past years, speech synthesis has almost matched natural speech (King and Karaiskos, 2009) in terms of intelligibility, when it comes to expressivity and spontaneity, TTS systems cannot yet achieve this goal. A major limitation arises from the fact that the textual surface form does not provide sufficient information for prosodic realization (Taylor, 2009), and that prosodic speech patterns are most commonly learnt within social interaction scenarios. This also leads to high degree of variability both for intra- and inter-speaker realizations.

However, prosody is regarded as an essential secondary communication channel (Huang et al., 2001) instinctually used by both the speaker, to encode his emotions and intentions, and by the listener, to aid his comprehension of the message. Therefore, when aiming at providing a natural human-computer interface, getting the message across is not sufficient, and as the human can encompass its psychological state of mind within speech, so should the responsive TTS machine provide an emotional feedback.

There are various manners through which the developers have tried to include additional layers of information within the front-end of speech synthesizers, and these are mostly related to including additional prosodic tags to the text to be synthesized. These tags can either be manually added (Carlsson et al., 2002) or automatically derived from text (Syrdal et al., 2001). But these are in most cases developed or available only for a limited set of languages (eg. English, French, Spanish etc.).

The speech processing resources and tools are still scarce for the Romanian language, and the lack of a common development framework makes it hard for researchers to compare results and make additional developments. However, during the last few years this situation has been slowly improving due to a number of shared or individual initiatives of research groups to make their tools and resources available. In terms of speech resources, the recent paper of (Stan et al., 2011) introduces a publicly available, high quality speech corpus named Romanian Speech Synthesis (RSS) database. This development enabled academic research in the area of corpora-based methods for speech synthesis for Romanian. The corpus consists of 3.5 hours of recordings divided into three sections: random newspaper section -- 1500 utterances (104 minutes); diphone coverage section -- 1000 utterances (53 minutes) and the fairy-tale section -- 1000 utterances (67 minutes).

The availability of the RSS corpus and the proficiency of statistical parametric speech synthesis systems in working with previously unseen patterns by successfully combining information from the available data has enabled our research to focus on the Romanian prosodic phenomena. We therefore present a prosodically-driven enhancement method of the RSS database through the addition of Tone and Break Indices (ToBI) (Silverman et al., 1992) style labels. As such, we called the new development the RSS-ToBI [1] corpus. The corpus is composed of a mixture of data obtained from RACAI's Natural Language Processing (NLP) Tools (Ion, 2007; Tufiș et al., 2008; Ștefănescu et al., 2012; Boroș et al., 2013) applied on the fairy-tale section of the RSS corpus, with an additional manually-created prosodic layer. The prosodic layer uses the ToBI standard for annotation with a series of adjustments introduced by (Jitcă et al., 2012) to suite the Romanian prosodic phenomena.

Having this resource at hand we then: (1) experiment with results obtained by embedding relevant prosodic information into the training phase of statistical parametric speech synthesis systems, (2) asses the performance of rule-based or statistical prosody

[1] The corpus is available through the META-SHARE platform: http://ws.racai.ro:9191

prediction methods and (3) compare user preference regarding basic synthesized speech and speech synthesized using manual and automatic ToBI labelling.
We cover two aspects regarding speech prosody: (1) determining an appropriate representation method and insertion of prosodic information in existing data (i.e. usually the training data used for TTS systems) (sections 2 and 3) and (2) automatic generation of prosodic information at runtime (section 4).

2. RSS-ToBI corpus description

There is still a level of disagreement regarding representation and description systems for prosody and there are several theories that support the existence of a prosodic hierarchy inside an utterance (Liberman and Prince 1977; Selkirk, 1984; Beckman and Pierrehumbert 1986; Nespor and Vogel 1983; Ladd 1996) and a number of description systems proposed for the task of prosodic labeling such as the International Transcription System for Intonation (INTSINT) (Hirst, 2000), the TILT intonation model (Taylor, 1998) or the Tones and Break Indices (Silverman, 1992). The later mentioned ToBI system is a widely accepted standard for prosodic annotation, which was initially designed to encompass the prosodic phenomena of English and was later adapted to other languages (e.g. the J–ToBI standard for Japanese (Campbell and Venditti, 1995) or the RoToBI standard for Romanian (Jitcă et al., 2012)).
The RSS-ToBI corpus is based on the prompts available in the fairy-tale section of the RSS dataset: 1000 utterances amounting to a total of 67 minutes of speech. The prompts were pre-processed using the RACAI NLP Tools to add typical local-context information required by TTS synthesis. This information includes: phonetic transcription, syllabification, stress prediction and part-of-speech (POS) tagging.
The prosodic annotation layer was built in two stages. In the first stage a number of 5 people were asked to listen and label the speech corpus, with the help of a custom designed visualization and editing tool that is compliant with the RACAI NLP Tools XML output format. Each annotator tagged the entire corpus. The initial inter-annotator agreement rate was below 40%. This result was to be expected, as the number of Romanian ToBI tags is large, as well as the fact that some tags describe similar patterns, and the annotators gave preference to one or a very similar another (e.g. H* and L+H*). Therefore a second evaluation stage was required. In it, a single speech expert went through the entire speech corpus, and, based on the primary annotations, manually edited and resolved the tags which had low inter-annotator agreement.
When grouping together both pitch accents and boundary tones, the corpus contains a total of 7022 labels (19 unique) (see Figure 1 for the complete set of labels and their occurrence within the corpus).

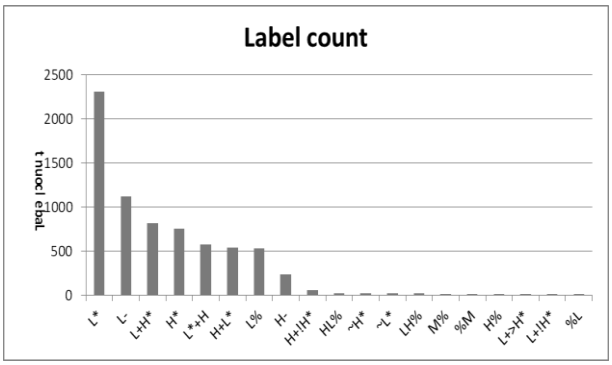

FIGURE 1 - TOBI LABELS OCCURRENCE WITHIN THE RSS-TOBI CORPUS

3. Evaluation of the RSS-ToBI Corpus

This section presents the tests performed to assess the usability and performance of the ToBI annotated corpus in a TTS system. Two HMM-based statistical parametric speech synthesis models were built using the ToBI labeled (system A) and the unlabeled (system B) versions of the RSS fairytale corpus. A number of 37 sentences were randomly selected and manually labeled from a previously unseen test set consisting of 19 news and 18 novel sentences. The test sentences were synthesized using both models and an anonymous preference test was conducted on a purpose-built website[2]. In the preference test, listeners were presented with speech samples from both systems, and, for each utterance they were asked to select from a lists of 5 preference options: (1) the systems sound identical, (2) system A sounds a little better than system B, (3) system A sounds much better than system B, (4) system B sounds a little better than system A and (5) system B sounds much better than system A. The participants were asked to carefully consider the prosodic aspect of the synthetic voices and to try and ignore the overall naturalness of the output.
The preference listening test is still ongoing, and we only present here the intermediary results. So far, we collected a number of 587 answers. In 52.81% cases, the RSS-ToBI labeled system was considered better than the unlabeled system, in 25.04% of the cases the systems were considered of equal quality and in 22.15% cases, the unlabeled system was considered better than the labeled one (see figure 2 for detailed results). It is important to note that the test respondents are not speech experts. Statistically, for a confidence level of 95% with a confidence interval of 5, we only needed a sample size of 377 answers.
Currently, for our 587 answers, with the worst-case 50% response distribution and a confidence level of 95%, we can be certain of the test's results with a confidence interval of 3.99%. This interval is sufficiently small to statistically prove that the ToBI labeled system is better than the unannotated system.

2 http://rslp.racai.ro/index.php?page=experiment/listening

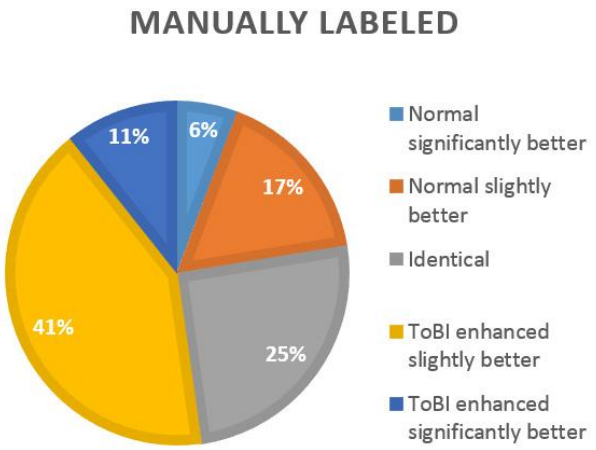

FIGURE 2 – MANUALLY LABELED SYSTEM PREFERENCE

4. Evaluation of the RSS-ToBI corpus prediction

To test the usability of the RSS-ToBI corpus, we also trained a number of well-known classifiers: Naive Bayes, J48, Random Forest, Support Vector Machine (SVM) and a Perceptron with the Margin Infused Relaxed Algorithm (MIRA) to perform automatic ToBI labeling on unseen data using features purely extracted from text. The feature set used in this experiment consisted of syllable n-grams, part-of-speech n-grams and the distances measured in syllables and words from the previously assigned label. The accuracies obtained using these classifiers with default parameters in a ten-fold cross validation procedure are: Naïve Bayes – 53.02%, J48 – 54.03%, Random Forest – 49.13%, SVM – 53.02% and MIRA – 54.32% (see table 1 for confusion matrix of the MIRA classifier).

We used the WEKA toolkit for every classifier (default parameters) except MIRA, for which we have an in-house implementation.

These results are to be expected, since at this point we only relied on a surface analysis of the text and did not use any natural rules/restrictions that would, for example, forbid the classifier to mistake an intermediate boundary tone for a final boundary or a pitch tone (e.g. the L- intermediate tone is systematically confused with the L% boundary tone and the L* pitch tone). Intuitively, advanced features extracted from the global context of the discourse can be used to enhance the results obtained by the data-driven labeling method, but this is a different topic

However, to answer the question whether a 54% ToBI-labeling prediction accuracy is sufficient to produce high quality speech, we used the automatically added ToBI labels to resynthesize the test data as follows:

- The system was trained on the original manually labeled speech corpus;
- The test data was re-labeled using the MIRA classifier and the initially trained system was used to resynthesize the newly labeled utterances;
- We used the same crowd-sourced platform and asked users to select their preference between the basic sentences (no prosodic information used during training and testing) and automatically labeled sentences (manual prosodic information provided during training and automatic labelling used on the test data).

The preference test yielded the results shown in figure 3. The results confirm our assumption that even automatically generated labels lead to better-sounding TTS: 32.5% think that the ToBI system is better, 42.1% make no difference and 25.4% think that the normal system is better.

TABLE 1 - CONFUSION MATRIX FOR THE MIRA CLASSIFIER

	L+H*	L*	H+!H*	L-	L%	H*	H+L*	L*+H	H-	~L*	~H*	LH%	HL%	%M
L+H*	**13.92**	79.75	0	2.53	0	3.8	0	0	0	0	0	0	0	0
L*	2.53	**94.51**	0	0.42	1.69	0.84	0	0	0	0	0	0	0	0
H+!H*	0	100	**0**	0	0	0	0	0	0	0	0	0	0	0
L-	0	9.62	0	**83.65**	6.73	0	0	0	0	0	0	0	0	0
L%	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0
H*	7.14	84.29	0	2.86	2.86	**2.86**	0	0	0	0	0	0	0	0
H+L*	10.64	82.98	0	2.13	4.26	0	**0**	0	0	0	0	0	0	0
L*+H	2.82	92.96	0	2.82	0	1.41	0	**0**	0	0	0	0	0	0
H-	0	30	0	65	5	0	0	0	**0**	0	0	0	0	0
~L*	0	100	0	0	0	0	0	0	0	**0**	0	0	0	0
~H*	0	100	0	0	0	0	0	0	0	0	**0**	0	0	0
LH%	0	0	0	0	100	0	0	0	0	0	0	**0**	0	0
HL%	0	0	0	0	100	0	0	0	0	0	0	0	**0**	0
%M	0	100	0	0	0	0	0	0	0	0	0	0	0	**0**

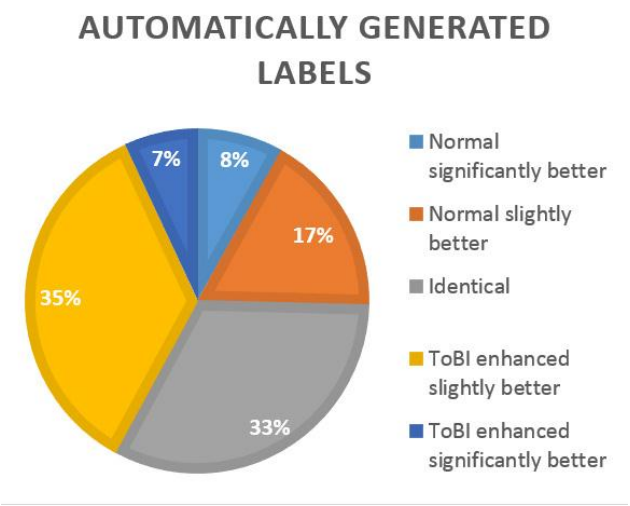

FIGURE 3 – AUTOMATICALLY LABELED SYSTEM PREFERENCE

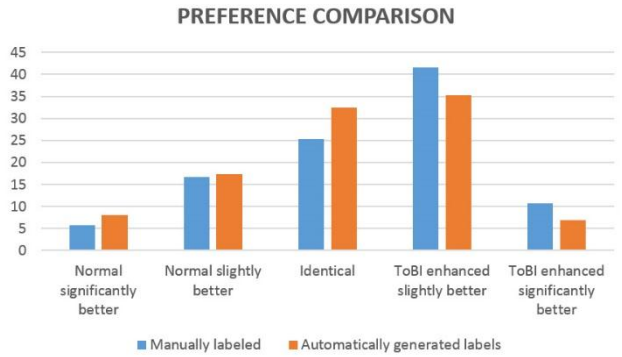

FIGURE 4 – PREFERENCE COMPARISON CHART

Statistically, due to the fact that we had fewer evaluators for this scenario (currently only 351, just shy of the 377 minimum sample size intended), for the same 50% response distribution (meaning that respondents have no bias towards any system) and a confidence level of 95%, we can be certain of this test's results with a confidence interval of 5.18%.

Figure 4 compares the two systems' evaluations side-by-side. The general response distribution is flatter for the automatic system than for the manual system, showing the effects of the automatic label generation errors. More respondents thought that the normal system sounds better and the ToBI system sounds worse in the automated label scenario compared to the same systems trained with manual labels. Also, 7.2% percent more thought that the systems sound identical in the automated-vs-manual comparison.

Though we had fewer respondents than in the manual labels preference test, the general consensus is similar: ToBI labels improve speech synthesis quality of TTS systems that train either on manual or on automatically generated labels.

5. Conclusions and future work

Smaller preference results were expected from the automatically labeled training data, given the accuracy of the classifier and its severe confusion between certain labels (see table 1) (e.g. L* is used instead of H* in 84% of the cases). Surprisingly, automatically labeled data was still considered better than the basic version. An explanation for this is that users preferred expressive speech over flat spoken utterances and the Romanian language offers freedom in speaking style. This does not mean that the underlying message is identical regardless of the pitch, tone and speaking tempo that are used but, for sentences spoken out of the blue, varying these voice parameters is sufficient to suggest naturalness.

The newly created speech corpus is a valuable asset to the Romanian language processing as it provides the means to train and test methods for automatically generating prosody directly from text. By analyzing the preference test results, it can be observed that in more than 50% of the cases the RSS-ToBI labeled system is preferred over the unlabeled one, while the unlabeled system is only preferred in about 20% of the cases.

One of the interesting aspects is that the statistical models for speech synthesis obtained using this corpus are suitable for distinct speaking styles (i.e. news and novel). The fact that the ToBI system consistently obtained better scores on both sections shows that it is possible to train and test a statistical parametric speech synthesis on different genres, provided that the prosodic annotations are performed according to the output needs. This enables researchers to test their own systems for automatic prosodic labels, provided that they map their labels onto the RSS-ToBI or they use and adapt this corpus as training data for their speech synthesis systems.

The corpus, as well as the other resources and tools needed to conduct a similar experiment are freely available for research purposes either through the META-SHARE platform[3], the Romanian TTS platform[4] or by contacting the authors.

Because of the promising results we obtained in the evaluation, we will focus future efforts at completing the manual labelling of the entire RSS corpus (not just the fairy-tale section) and we will continue our research on creating better automatic methods for prosodic labelling of Romanian text, that rely on more than just surface features and include information extracted from the global context of the text.

[3] http://ws.racai.ro:9191

[4] http://romaniantts.com/new/rssdb/rssdb.php

6. References

Boroş, T., Ştefănescu, D. and Ion, R. (2013). Handling Two Difficult Challenges for Text-to-Speech Synthesis Systems: Out-of-Vocabulary Words and Prosody -- A Case Study in Romanian. Book chapter in Where Humans Meet Machines, ISBN 978-1-4614-6934-6, pp. 137—161 edited by Amy Neustein, Springer.

Carlsson, J., Paiz, C., Wolff. K., Nordin, P. (2002). *Interactive Evolution of Speech using VoiceXML Speaking to you GP System*. In Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics, VI, page 58--62. IIIS.

Huang, X., Acero, A., & Hon, H. W. (2001). *Spoken language processing*. New Jersey: Prentice Hall PTR.

Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD thesis (in Romanian). Romanian Academy, Bucharest.

Jitcă, D., Apopei, V., & Păduraru, O. (2012). The Ro-Tobi Annotation System and the Functional Analysis Perspective Of The Romanian Intonation. In Proceedings of CONSILR, ISSN 1843-911x, 3.

King, S. and Karaiskos, V. (2009). *The Blizzard Challenge 2009*. In Proc. Blizzard Challenge Workshop, Edinburgh, UK.

Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992). *TOBI: a standard for labeling English prosody*. In Proceedings of ICSLP (Vol. 2, pp. 867-870).

Stan, A., Yamagishi, J., King, S., & Aylett, M. (2011). The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate. In Speech Communication, 53(3), 442-450.

Ştefănescu, D., Ion, R., & Hunsicker, S. (2012). *Hybrid parallel sentence mining from comparable corpora*. In Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012), pp. 137—144, Trento, Italy.

Syrdal, A. K., Hirschberg, J., McGory, J. T., Beckman, M. E. (2001). Automatic ToBI prediction and alignment to speed manual labeling of prosody, Speech Communication 33(1-2):135-151.

Taylor, P. (2009), Text-to-speech synthesis, Cambridge University Press.

Tufiş, D., Ion, R., Ceauşu, A., & Ştefănescu, D. (2008). RACAI's Linguistic Web Services. In Proceedings of the 6th Language Resources and Evaluation Conference-LREC.