Automatic phonologial phrase and prosodic event detection for the extraction of syntactic and semantic/pragmatic information from speech

NKFIH PD-112598

Closing report

György Szaszák, PI

Dept. For Telecommunications and Media Informatics,
Budapest University of Technology and Economics

2017, December

1. Scope

As we have outlined in the research plan, speech prosody is one of the most significant and basic building blocks of spoken language and serves to carry information about many important aspects such as syntax and speech segmentation, discourse function, salience, speaker attitude and emotions, (Jurafsky and Martin, 2008; O Shaughnessy, 2007) etc.

The lack of proper prosody can make the speech sound unnatural and hard to follow, even though it might be fully intelligible. This explains why prosody generation modules have so crucial role in text-to-speech (TTS) synthesis systems since the beginning (Dutoit, 1997), especially in today's state-of-the-art TTS systems, whose focus is set not only on sounding natural, but also on expressing emotions and attitudes (Bulut and Narayanan, 2002).

On the other side, prosody can also be exploited related to Automatic Speech Recognition (ASR), and the present research project targeted primarily laying down a research basis for such kinds of future applications. Most ASR systems treat speech as a word sequence, and then, based on acoustic models (e.g. phonemes) and a language model (typically an N-gram), a so-called recognition network is created, along which the speech frames are aligned using the Viterbi-algorithm. The recognition hypothesis is yield by the most likely alignment path in the network, given the acoustic observation (e.g. speech frames). This processing chain does not incorporate prosody, which, due to its supra-segmental nature is hard to treat in a system based on segmental level processing.

Recently, ASR technology has seen several higher level applications built on top of pure speech-to-text output. Speech understanding for information extraction, question answering and slot-filling, spoken term detection, speech-to-speech translation (Aguero et al., 2006) all require syntactic and semantic processing of the produced word chain hypotheses, tasks where prosody can and should be exploited to improve performance. Auxiliary tools coupled with ASR also can benefit from prosody

in emotion detection, training and diagnosis for speech disorders, dialogue modeling, etc. A straightforward way prosody can be used in these tasks is to provide a segmentation of the speech stream (Shriberg et al., 2000; Vicsi-Szaszák, 2010), which reflects the information structure (Selkirk, 2001; Gallwitz et al., 2002; Szaszák-Beke, 2012), focus, emphasis and modality (Král et al., 2005), etc. Providing capitalization and punctuation for the ASR output itself is also a task where prosody can be efficiently exploited.

The novelty of the current research was that it employed a phonological phrase level approach by a coherent modeling of speech prosody. Far the most common approach in prosody processing is to look for event markers such as F0 accents, boundary tones etc. (Silverman, 1992). In contrast, the approach proposed by Vicsi and Szaszák (2010) treats prosody as an in itself meaningful entity and allows for recovering the complete prosodic structure at once (Szaszák-Beke, 2012) instead of just sampling parts of it driven by the actual task. The main goal of the research within the project was to extend, optimize and enhance this modeling framework by (i) extending it to other languages, (ii) show that this approach is applicable to non fixed-stress languages, (iii) optimize the algorithm in terms of acoustic features and (iv) make it as language independent as possible, (v) develop and evaluate the elaborated framework in tasks where prosody processing can be exploited: automatic phrasing and stress detection, automated corpus labeling, separation of syntactically motivated stress and automatic punctuation of ASR output.

In this report, the tasks defined in the research plan are overviewed one-be-one to show the outcome of each task 2.1-2.5. Task 2.6 is an additional task, undertook when applying for a 3 months prolongation of research funding. Tasks 2.7 and 2.8 were not explicitly planned, but emerged as a possible exploitation and enhancement of the planned research and hence worth investigating. In the following we provide a brief overview of the research activity and its results.

2. Research results

2.1. Improve the Hungarian phonological phrase detection approach and explore adaptation possibilities for spontaneous speech

For Hungarian, a phonological phrase detection approach has been developed (Vicsi and Szaszák, 2010; Szaszák and Beke, 2012). This approach relies on the detection of prominence or stress and also involves the classification of the intonation contour, using F0 and energy features. Phonological phrases (PP) constitute a prosodic unit, characterized by a single stress and some preceding/following intonation contour. As this contour is specific, phonological phrases can be classified and hence modeled separately, in a data-driven machine learning approach. The distinction between phonological phrases consists of two components: the strength of stress the phonological phrase carries and the intonation contour. In this way 7 different types are distinguished for Hungarian. Modeling is done with HMM/GMM models, and phonological phrase segmentation is carried out as a Viterbi-alignment of phonological phrases for the utterances requiring segmentation. The baseline system implementing this approach was available for Hungarian at the start of the research and served as the starting point of current research.

The research plan foresaw the improvement of the detection accuracy by considering probabilistic features (Garner et al., 2013) in task 2.1. Probabilistic features were preferred because they allowed for a much more flexible and efficient approach.

In the baseline system, fundamental frequency (F0) and wide-band energy (E) are used as acoustic features. First we focused on the F0 component, as F0 interpolation is important as recent research revealed that a continuous, overall defined F0 feature stream yields better results in prosody processing, at least in read, formal speech (c.f. Yamagishi et al, 2009; Yu-Young, 2011). Normally, F0 is defined only for voiced frames of the digital speech signal. This can be made continuous by interpolating F0 values over the unvoiced frames, which is a commonly used technics in F0 processing. Our research compared two interpolation technics to the uninterpolated case: a technic applying overall interpolation and a technic applying partial interpolation based on constraints related to F0 change and the duration of the voiceless speech segment. Interpolation is made linearly, but in logarithmic domain, or alternatively, is extracted form a Viterbi-like constrained extractor as implemented in Kaldi (extract-kaldi-pitch-feats) (Ghahremani et al., 2014).

The three compared scenarios were as follows:

- Use the F0 contour as produced by a conventional pitch tracker (Snack V2.2.10 in our case, doubling/halving errors corrected automatically);
- Use a continuous F0 contour, interpolated at all unvoiced parts;
- Use a partially F0 interpolated contour, where interpolation is omitted if the length of the unvoiced interval exceeds a limit (250 ms in the experiments) or if F0 starts significantly higher (suspected pitch reset) than it was before the unvoiced segment (criterion applied in the experiments: $F0\_former * 1.1 < F0\_current$).

The motivation to constrain the disruption of the F0 contour in the partial interpolation scenario comes from the following considerations: (i) the silence limit is set because a silent period longer than 250 ms can hardly be considered as fluent speech. In such cases the speaker may not employ an F0 reset as the silence in itself can be a clear acoustic marker of phrase boundary. If this happens interpolation may mask the phonological phrase boundary, although energy features are still likely

to signal it. (ii) Medium strength pitch resets may often be smoothed by the F0 interpolation, which makes further detection more difficult. Hence, a factor of 1.1 is preferred in order to avoid that microprosodic disturbances give false phonological phrase boundary detection, which may happen in the vicinity of long plosives for example.

Results showed (Szaszák-Beke, 2015; Szaszák-Tulics-Tündik, 2014) that formal and informal speech behave differently: in formal speech, the overall interpolation is the best solution (~15% relative improvement compared to baseline), whereas in informal speech, a constrained, partial interpolation outperforms the overall one (by ~5% relative).

Probabilistic features add a confidence like measure of voicedness to the F0 stream, beside usually interpolating it over the whole utterance. Regarding probabilistic features, a modest improvement (~1%) was observed upon the improvement already achieved by improving the F0 interpolation.

Overall, the phonological phrase detection and segmentation system has reached close to 90% precision and recall in Hungarian by allowing one syllable long (~250 ms) tolerance in boundary detection.

2.2. Research on automatic duration feature extraction methods

Prosodic feature extraction is easy for energy, somewhat more problematic for F0 (see the previous section 2.1), but for duration extraction, the underlying segmental level phone boundaries are required. In order to maintain the fully automatic operation, we considered two alternatives: (i) using an ASR for transcription and forced alignment, or (ii) using a broad phoneme classifier.

For the experiments we used ASR monophone models with equal probability phoneme unigram language model trained for the respective language. Generally, if we regard phonological phrase segmentation as an auxiliary information source in automatic speech understanding applications, keeping its computational requirement low is beneficial in order to let the overall system concentrating its effort for resource demanding tasks.

As presented already, the baseline phonological phrase detection/segmentation system relies on fundamental frequency (F0) and mean energy (E) as principal acoustic features. Beside, the used Hidden Markov Model (HMM) architecture is already capable of handling durational patterns in terms of the dynamic time warping resulting from the application of a Viterbi-decoding as part of the segmentation. In addition, research was interested in adding explicitly durational features to the feature space. From the obtained phone boundaries (yielded by ASR), two types of durational feature streams are created: one reflecting the length of the vowels, and another one reflecting the length of syllables. Both streams are smoothed with a wide span mean filter in order to obtain an overall continuous feature stream for the the whole utterance. Another advantage of the mean filtering is the removal of sudden changes in the signal value. No normalization for phone types is performed. With the feature streams appended separately and together to the baseline feature set, no improvement was seen either for Hungarian or for French.

Summarizing the outcome, for Hungarian and French we hypothesized that as these languages are fixed stressed, duration played a minor role in PP segmentation – or, at most, they provide only redundancy as F0 and energy already reflect the phrasing information. Our results confirmed this hypothesis experimentally (adding duration features did not lead to significant improvement in performance).

For English and German, which use lexical stress, we hypothesized that duration features may help PP segmentation as these languages, especially English, are considered as languages where duration in prosody plays a more important role in stress. Nevertheless, with respect to PP segmentation, we were not able to confirm this hypothesis either for English or German. We explain this as follows: given the HMM structure already implicitly involves some durational modeling, these results confirm that in the used framework, adding durational features is unproductive with respect to phonological phrase segmentation. For the same reason, these results cannot be interpreted as contradictory to theory, that hypothesizes that durational lengthening is a phonological phrase boundary marker. In our case, the flexibility of the HMM framework allows for a dynamic time warping of the patterns, hence may implicitly exploit durational knowledge. This explains why adding these features explicitly does not lead to any further improvement.

2.3. Research on phonological phrase detection in other languages

The research plan targeted the adaptation of the Hungarian system for French, English and German. The basic interest here was to analyze whether the phonological phrase detection approach elaborated for Hungarian is adaptable for these languages. This work involved prosodic labeling of phonological phrases for at least one hour of data from different speakers in each language. Prior to this labeling, the set of phonological phrases had to be defined for each language. The set was iteratively modified as manual labeling was progressing. As data set we chose the SIWIS database (Garner et al., 2014), kindly made available for our research by Idiap Reserach Institute in Switzerland, which is a parallel French, German, Italian and English spoken corpus. Manual labeling and annotation of this corpus was mostly carried out by a technical assistant and university students, supervised by the principal investigator.

According to the work plan, we implemented phonological phrase alignment for French, English and German. Unlike Hungarian and French, English and German have lexical stress, despite still showing a tendency for marking the left edge of the phonological phrases, similar to Hungarian. However, if lexical constraints require, stress is moved from the first syllable. To cope with these differences, data processing (in the second year of the project) was planned such that phrases with left edge stress are separated form phrases where stress is moved from the first syllable. When stress detection is addressed, the prominent syllable have to be identified within the phrase, as we cannot rely on phrase edges. Fortunately, this is easily feasible, as a phonological phrase by definition contains a single prominence, hence a maximum search on the F0 contour is sufficient.

Regarding the performance of the phonological phrase alignment, we experience a slight drop in precision and recall when switching to a language with lexical stress, resulting from the presence of lexical stress and the different organization of the Germanic languages in terms of prosody compared to Hungarian or French. However, precision and recall are high enough for promising exploitation in further speech technology applications: PP boundary recovery by allowing 250 ms deviation (~ one syllable long) from the reference one are as follows for the operating points characterized by equal precision and recall of the monolingual systems:

- Hungarian: 89.4%
- French: 89.0%
- English: 83.4%
- German: 77.8%

Our hypothesis that the PP modeling and segmentation framework is adaptable for other languages has hence been experimentally confirmed, as precision and recall of phrase boundary placement is

comparable to human inter-annotator agreements, typically around 80% in prosodic labeling tasks (Pitrelli et al., 1994).

2.4-5. Research on the syntax-prosody interface and salient event detection

For Hungarian, partial recovery of the syntactic structure was possible from the pure speech signal, based on phonological phrase detection (Szaszák and Beke, 2012). We hypothesized that prosodic stress may result from either syntactic or semantic /pragmatic effects and that by separating the two, further semantic meaning could be extracted, which is beyond the syntactic functions (i. e. reflecting the information structure) of speech prosody. In this section we present 2.4 and 2.5 research tasks together, as they are closely related.

Based on acoustics, both syntax and semantics triggered stress have similar characteristics and hence, separating them based purely on acoustics has its limits: some PP types show a tendency to be aligned by relatively low confidences to syntactic units. Therefore, we preferred an approach where text analysis is also carried out. Based on this, syntactic phrasing can be obtained, which can then be compared to the phonological phrasing.

In a simpler labeling scheme, native, non expert Hungarian annotators were asked to read written sentences and mark the words they would expect to have any kind of notable acoustic prominence if the sentence were read aloud: each of them received a response sheet with all of the 1948 sentences used in the experiment and was instructed to mark the words (s)he thinks should be stressed. For each sentence, it was allowed to mark as many words as desired. As Hungarian is a fixed stress language (stress is assigned to the first syllable of a word), a pure marking of the word expected to carry the prominence is sufficient. There is no distinction regarding the strength of prominence, but a simple binary decision is made (prominent or not prominent). This labeling was carried out by a single male annotator for the whole corpus, accompanied by two others (1 male, 1 female) on a smaller subset (on 10 % of randomly chosen data) to evaluate inter-annotator agreement. This latter was found 86.2%.

Prosody based prominence marking is carried out by the automatic PP alignment tool developed within the project, which operates purely on the acoustic speech signal (relying mainly on F0 and overall energy as acoustic features) and does not have any linguistic input. Performing PP segmentation yields the stressed syllables as well.

Comparing the two prominence markings we can obtain the set of prominent syllables which are not marked based on text (syntax) and hence are likely to be of semantic or pragmatic origins. Supposing that syntax is the main governing factor in text based prominence marking, we observed that

- syntax often (20.5%) marks alone the prominence without the intervention of prosody;
- prosody often marks prominence that is unavailable in human syntax motivated prominence annotations (48.9%), although the eventual role of prosody in signaling word boundaries is an alternative hypothesis here. However, the fact that prosody accounts for almost half of the prominence markings alone and that automatic word boundary recovery based on prosody could not reach more than 78% precision (Vicsi-Szaszák, 2010) suggests that prosody has a function in human speech which is beyond syntax;
- often (30.6%) both syntax and prosody mark a prominence. This does not necessarily justifies a deduction like “prosody is there to reflect syntax and hence the information structure”, but may also mean that prosody further strengthens an already syntactically

signaled emphasis.

The observed sharp differences (Szaszák-Beke, 2017) between text and prosody based prominence markings were beyond our expectations (overlap is less than 1/3 of all cases). From our research perspective which is purely practical and cannot prove or disprove any linguistic theory, the obtained results confirm our research hypothesis that prominence predictions based on acoustics and syntax are significantly different in the used evaluation setup. This is an important outcome considering automatic speech recognition, speech-to-speech translation or speech understanding applications, where it seems to be crucial to include both acoustic and textual analysis if content analysis or extraction of the meaning is targeted. For example, in speech-to-speech translation, a pure text based regeneration of the prominence on the target language side would result in loosing all prominence marked only by prosody and hence the information conveyed in this manner. These findings favor a theoretical approach of prominence, whereby prosody is not only an auxiliary player subordinated to syntax, but has rather an individual role in further structuring of the information and providing cues for perception orientation and for the decoding of paraverbal information. The role of prosody may also include some “rescoring” or re-weighting of the syntactically marked emphasis or information structure in general. Indeed, this is the difference which makes human speech richer than pure text.

Beside the basic research presented so far in this subsection, the PP segmentation framework was used for the automatic labeling of a TTS inventory (Szaszák et al., 2015). During the preparation of a speech corpus for TTS purposes, a basic step is the precise labeling of prosodic stress. This can be done manually or automatically. Manual stress annotation is relatively precise, however, it is very time consuming and introduces subjectivity. For example, inter-annotator agreement scores using ToBI annotation are usually found between 70-80% for pitch-accents in English (Pitrelli et al., 1994). Our personal experience shows that syntax influences human annotation highly, even if done by high-qualified experts. Our earlier results presented in this section suggest that if stress is syntactically predictable, its marking may be missed by prosodic features. An alternative way of stress labeling is automatic, and based on the transcription of the speech material. This approach can also lead to a corpus suitable for good quality synthesis, however, it also suffers from errors resulting from the mismatch between syntactic and prosodic marking of stress. If stress marking is done in a rule-based manner as is often the case, further difficulty evolves when stress prediction fails, due to the well known lack of generalization capabilities of the rule-based approach. However, this latter problem can be cured, if speech corpora for TTS are well planned and read carefully, hence utterances requiring special attention can be handled by exceptions.

Given these difficulties of stress labeling, and the fact that even human labeling is somewhat ambiguous (inter-annotator agreements up to 80%), an approach predicting stress based on audio could be used to obtain stress labeling. Given that PP segmentation can be used for stress detection, labeling is straightforward using the produced PP segmentation framework. TTS results (mean opinion scores of users rating the TTS as well as decision tree analysis of the parametric TTS engine showed superior performance compared to baseline exploiting human stress annotation (Szaszák et al., 2015).

2.6. Punctuation based on phonological phrasing

As the PP segmentation framework became more mature, our research also focused on possible future applications. One of these is automatic prediction of punctuation marks. Our project was initially planned for 36 months, but part of the funds could be spared and hence, we applied for a 3 months prolongation by undertaking a new task: implement and evaluate punctuation prediction for unpunctuated text.

Inserting punctuation marks into the word sequence hypothesis produced by ASR systems has long been neglected, as research was mostly concerned by reducing word error rates and augmenting transcription accuracy for the word chain on one hand, and punctuation is not relevant in applications where the text output is not directly required, but rather the system is expected to react according to the received commands or queries, on the other hand. In dictation systems, where punctuation is the most relevant, a telegraphic style explicit dictation of the punctuation marks was foreseen, similarly to commands intended to provide text formatting, i.e. "*SET_BOLD SET_INITIAL_CAPITALS dear mister smith COMMA SET_NORMAL NEWLINE ...*". Nevertheless, providing punctuation automatically is the only applicable approach in several use-cases, i.e. for closed captioning of audio data (subtitling), transcription of meeting records, audio indexing followed by text analysis etc. In dictation systems, also, it is more natural and easier to speak normally, whereby the system automatically detects where punctuation is necessary.

Speech prosody is often used for punctuation (Batista et al., 2012; Tilk-Alumae 2017) as prosody is known to reflect the information structure of the speech to some extent. Features representing intonation, stress and pausing (F0 slopes and trends, pause durations) are found to be the most efficient. Prosody based approaches have the advantage of being independent of ASR errors, albeit usually still yield weaker performance than text based punctuation approaches (Tündik-Tarján-Szaszák, 2017). When lightweight models and real-time operation are required (closed captioning, on-the-fly transcription, subtitling, speech-to-text translation etc.), however, prosody is the most powerful way to predict punctuation.

The proposed punctuation model (Moró-Szaszák, 2017) exploits the expected correlation between phonological phrasing and punctuation marks. As the phonological phrasing represents the building blocks of sentence level intonation, we model them as a sequence and map this sequence to the sequence of the punctuation marks. The most suitable machine learning framework for such tasks is using recurrent neural networks with Long-Short Term Memory cells (LSTM). We start from the automatic PP alignment and the word sequence, which are supposed to be known (as PP sequence hypothesis from the PP segmentation and word sequence hypothesis from ASR). PPs are constrained to start and end on word boundaries, as punctuation marks may also be required at word boundaries (so called *slots*). Then, we extract the following features to be input to the RNN:

- the type of the PP.
- the duration of the PP.
- the duration of short pause or silence following the PP.

Overall punctuation results with this lightweight system were up to 83% precision and 45% recall for commas, 82% precision and 89% recall for periods on a read Hungarian dataset. These numbers are comparable to state-of-the-art performance of such systems reported so far (Batista et al., 2012; Tilk-Alumae, 2017).

It is an interesting question how the users themselves perceive punctuation accuracy and quality. All the measures used so far are objective measures and these are computed from comparing the automatic punctuation to the reference one. Taking as an example the closed captioning of live video or audio with ASR, from a user perception point-of-view, subtitles are visible for some seconds, whereas the user concentrates on getting the meaning and following the video as well. In other words, it is more important, what is written, than how it is written. In addition, we may suppose that an unconscious error repair mechanism is functioning, which, just like in self repairing coding, restores the correct punctuation sequence or ignores the errors in it, as far as error ratios are below a critical threshold. We adopted this paradigm of the cognitive infocommunication (Baranyi-Csapó-Sallai, 2015) for our subjective tests.

To carry out the subjective tests, we select 4 samples, composed of 5-7 coherent sentences from the BN corpus, and prepare 3 types of text for each:

- a reference transcript with automatic punctuation (AP);
- an ASR transcript with reference punctuation (AT);
- reference text with reference punctuation as a control set (CTRL).

Users were asked to rate the text on a scale from 1 to 5 according to the following guideline: "*In the following text word or punctuation errors may appear. To what extent do these errors influence your ease of understanding? "*. During the evaluation, we contrasted AP with CTRL and AT with CTRL.

35 subjects, 28 male and 7 female with 29,6 years mean age took part in the tests, assessing two types of text out of the three possible. Most of them were university students academic sector employees. The subjects got the texts on a sheet and they had to read through once the 2 short blocks. One of the blocks tested for word errors, the other one for punctuation errors. The users were unaware of whether they receive a correct (reference or 100% accurate automatic) text or an incorrect text with eventual errors. They had to rate the texts according to how disturbing the errors were regarding the interpretation of the meaning (with score 5 = not disturbing at all to score 1 = text not understandable due to the errors).

On the ratings we calculated Mean Opinion Score (MOS) and performed a chi-square test to see whether differences are statistically significant. Surprisingly, MOS for AP (4.28) is higher than for the control blocks (4.19), but it is more important that even by 1% significance level ($p<.01$), subjects were not able to make a difference between correct and erroneous texts in terms of punctuation. Spotting ASR errors is easier, regarding the AT vs. CTRL task we found a statistically significant difference in ratings by 5% significance level ($p>.05$). MOS for AT was 4.05.

2.6. A physiological-phonological hybrid approach

Although not explicitly planned for the project, a new algorithm based on the physiological pitch production model has been proposed (Honnet, P.E., Gerazov, B. and Garner, P.N., 2015). The algorithm is called weighted correlation atom decomposition (WCAD). The idea it to use atomic building elements for the pitch contour of the utterance, and try to match these by minimizing some error criterion. We had the opportunity to cooperate directly with the developers of this new method. Given this algorithm was promising for the current task, we evaluated it for phonological phrase segmentation and compared results to our baseline (Szaszák et al., 2016). Results show that our baseline performs better, but combining the two systems into a hybrid yields some improvement in relevant operation points for phonological phrase detection when better than 250 ms time accuracy is desired, as detection time accuracy has also been improved with the hybrid system. The principal investigator involved a PhD student for this task, and this work – beyond of our original research plan – has been evaluated in international collaboration, which hopefully increases the visibility of our original research in the international scientific scene. Beside Hungarian, experiments were extended to the French language as well.

2.7. Speech summarization

A possible approach of summarizing written text is to extract important sentences from a document based on keywords or cue phrases. Automatic sentence segmentation (tokenization) is crucial before such a sentence based extractive summarization. The difficulty comes not only from ASR errors, but also from missing punctuation marks, which are fundamental in syntactic parsing and POS tagging (disambiguation). A prosody based automatic tokenizer inspired by PP segmentation was proposed (Beke-Szaszák, 2016) to recover intonational phrases (IP). IPs can be used as sentence like units in further analysis. Summarization was compared to a baseline version using tokens available from human annotation . The baseline tokenization relies on acoustic (silence) and syntactic-semantic (sintactically or semantically closely together belonging) axes.

The PP segmentation is conceived in such a manner that it encodes upper level IP constraints (as IP starter and ending PPs, as well as silence are modelled separately), and hence is de facto capable of yielding an IP segmentation, capturing silence, but also silence markers (often not physically realized as real silence). We used the IP tokenizer in an operating point with high precision (96% on read speech) and lower recall (80% on read speech) to obtain sentence like IP units for further processing. We consider less problematic missing a token boundary (merge 2 sentences) than inserting false ones (splitting the sentence into 2 parts).

An important outcome of the experiments is that the automatic, IP detection based prosodic tokenization gave almost the same performance as the human annotation based one (in soft comparison it is even better). The overall best results were 62% recall and 79% precision (F 1 = 0.68). Subjective rating of the summaries gave 3.2 mean opinion score.

Conclusions

Although this research report is in itself a summary of 39 months' effort, summarizing the results of the project in bullet points is intended to help the assessment procedure:

- The phonological phrasing framework was optimized, and also compared to an alternative approach in standalone and hybrid scenarios;
- The framework was successfully extended for new languages, including languages with lexical stress;
- An algorithm suitable for separating syntactically motivated prominence from non syntactically motivated prominence was proposed and evaluated;
- The framework was successfully adopted and used in cases where TTS and ASR systems can directly benefit from the improvement or speed-up yield by phonological phrasing: automatic labeling, stress detection, punctuation prediction, healthcare assessment (mental and speech disorders) and speech summarization.

Dissemination and publication of results

We were active in several national and international **conferences** to publish our latest results. An exhaustive list of all publications can be found in the respective section of the complete report. Regarding journal publications, by the time of the closing report, we had 2 of our **international peer-reviewed journal** publications and 3 others (peer-reviewed also) in the LNCS series published. We have another one in review (Speech Communication, Elsevier), another submitted (Acta Polytechnica Hungarica) and another written and to be submitted to a special issue upon invitation of Speech Computer and Language closing on the 25th January, 2018. We hope that many of them will be accepted for publication, unfortunately review processes are rather long, and considering that research results come to a maturity towards the ending of the project, our best effort was to write and submit them, we are looking forward to keep the NKFIH up-to-date on the status of these if required.

We would like to highlight also that the principal investigator gave a **plenary talk** summarizing the results of this project at IEEE CogInfoCom 2017 International Conference titled: „Speech Prosody: a Barely Known, yet Surprisingly Versatile Cue and its Rich Exploitation Possibilities in Cognitive Infocommunications” (http://www.coginfocom.hu/conference/CogInfoCom17/plenary.html).

We would like to highlight as well that on the same conference IEEE CogInfoCom 2017 International Conference our paper „*A prosody inspired RNN approach for punctuation of machine produced speech transcripts to improve human readability*” by Anna Moró and György Szaszák won the **best paper award** of the Program Committee.

Acknowledgements

The principal investigator and all project participants once again express their gratitude to NKFIH for funding this research, and thank the review and assessment committees for their work and effort.

References

Aguero, P.; Adell, J. and Bonafonte, A.: Prosody generation for speech-to-speech translation. In Proc. of ICASSP 2006, volume 1, pages 700–705, 2006.

Baranyi, P., Csapo, A. and Sallai, G.: Cognitive Infocommunications (CogInfoCom). Springer, 2015.

Batista, F., Moniz, H., Trancoso, I. and Mamede, N.: Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts, Transactions on Audio, Speech and Language Processing, 20(2), pp. 474–485, 2012.

Beke, A., Szaszák, G.: Automatic summarization of highly spontaneous speech. LECTURE NOTES IN COMPUTER SCIENCE 9811: pp. 140-147. (2016)

Bulut, M. and Narayanan, S.: Expressive speech synthesis using a concatenative synthesizer. In ICSLP 2002, Denver, Colorado, USA, September 2002.

Gallwitz, F., Niemann, H., Nöth, E., Warnke, W.: Integrated recognition of words and prosodic phrase boundaries. Speech Communication, Vol. 36. pp. 81-95. 2002.

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

Garner, P.N., Cernak, M. and Motlicek, P.: A simple continuous pitch estimation algorithm. IEEE Signal Processing Letters, 20(1):102–105, January 2013.

Garner, P.N., Clark, R., Goldman, J.P., Honnet, P.E., Ivanova, M., Lazaridis, A., Liang, H., Pfister, B., Ribeiro, M.S., Wehrli, E. and Yamagishi, J.: Translation and prosody in Swiss languages. In *Nouveaux cahiers de linguistique française* (No. EPFL-CONF-199815). 2014.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J. and Khudanpur, S.: A pitch extraction algorithm tuned for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (pp. 2494-2498). 2014

Honnet, P.E., Gerazov, B. and Garner, P.N., 2015, April. Atom decomposition-based intonation modelling. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 4744-4748), 2015.

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

Koehn, P. and H. Hoang, H.: 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), page 868876, 2007.

Král, P., Klečková, J. Cerisara C.: Sentence Modality Recognition in French based on Prosody. Proceedings of World Academy of Science, Engineering and Technology, Vol. 8, October 2005. ISSN 1307-6884. pp 185-188.

Moró, A. and Szaszák, G.: A Phonological Phrase Sequence Modelling Approach for Resource Efficient and Robust Real-Time Punctuation Recovery In: Proceedings of Interspeech: Situated interaction. Stockholm, Sweden. Causal Productions, pp. 558-562. 2017.

O'Shaughnessy, D.: Modern methods of speech synthesis. IEEE Circuits and Systems Magazine, 7(3):6–23, 2007.

Pitrelli, J.F., Beckman, M. E. and Hirschberg, J.: Evaluation of prosodic transcription labeling reliability in the ToBI framework," in Proceedings of the 1994 International Conference on Spoken Language Processing, vol. 1, pp. 123–126, 1994.

Shriberg, E., Stolcke, A., Hakkani-Tür, D. and Tür, G.: Prosody-based automatic segmentation of speech into sentences and topics. Speech Communication, vol. 32, no. 1-2, pp. 127-154, 2000.

Silverman, K. et al.: ToBI: A standard for labeling English prosody. In *Second International Conference on Spoken Language Processing*. 1992.

Szaszak, G., Nagy, K. and Beke, A.: Analysing the correspondence between automatic prosodic segmentation and syntactic structure. In: INTERSPEECH-2011 Conference Proceedings: Speech Science and Technology for Real Life. Firenze, Italy, ISCA, pp. 1057-1060. 2011

Szaszák, G. and Beke A.: Exploiting Prosody for Syntactic Analysis in Automatic Speech Understanding. JOURNAL OF LANGUAGE MODELLING 2012:(1) pp. 143-172. (2012)

Szaszák. G. and Beke, A.: Toward Exploring the Role of Disfluencies from an Acoustic Point of View: a New Aspect of (Dis)continuous Speech Prosody Modelling. LECTURE NOTES IN COMPUTER SCIENCE 9302: pp. 369-377. (2015)

Szaszák, G., Beke, A., Olaszy, G, and Tóth B.P.: Using automatic stress extraction from audio for improved prosody modeling in speech synthesis. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2015). pp. 2227-2231, 2015.

Szaszák, G. and Beke, A.: An empirical approach for comparing syntax and pros ody driven prominence marking. PHONETICIAN 114:(1) pp. 46-57. (2017)

Szaszák, G., Tulics, M.G., Tündik, M.Á.: Analyzing F0 Discontinuity for Speech Prosody Enhancement. ACTA UNIVERSITATIS SAPIENTIAE ELECTRICAL AND MECHANICAL ENGINEERING 6: pp. 59-67. (2014)

Szaszák, G., Tündik, M.Á., Gerazov, B., Gjoreski, A.: Combining atom decomposition of the F0 track and HMM-based phonological phrase modelling for robust stress detection in speech LECTURE NOTES IN COMPUTER SCIENCE 9811: pp. 165-173. (2016)

Tilk, O. and Alumäe, T., 2016. Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In *INTERSPEECH* pp. 3047-3051 (2017).

Tündik, M. Á., Tarján, B., Szaszák, G.: Low Latency MaxEnt- and RNN-Based Word Sequence Models for Punctuation Restoration of Closed Caption Data In: Camelin, N., Estève, Y. and Martín-Vide, C. (eds.): Statistical Language and Speech Processing. Springer, pp. 155-166. 2017.

Vicsi, K. and Szaszák, Gy.: Using prosody to improve automatic speech recognition. SPEECH COMMUNICATION 52:(5) pp. 413-426. (2010)

Yamagishi, J; et al.: A robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Transactions on Audio, Speech and Language Processing, 17(6): 365–375, 2009.

Yu, K. and Young, S.: Continuous F0 modeling for HMM based statistical parametric speech synthesis. IEEE Transactions on Audio, Speech and Language Processing, 19(5):10711079, 2011.