Automatic punctuation recovery in read and spontaneous Hungarian using a recurrent neural network based sequential model for phonological phrases

György Szaszák and Anna Moró
Budapest University of Technology and Economics

Despite the advances in automatic speech recognition technology, the automatic placement of punctuation marks is still an open issue; in most dictation systems, users have to explicitly dictate commas, sentence terminal punctuation marks, etc. In some applications – such as office or medical dictation (Vicsi et al., 2006), – this is an unnatural, but possible way of adding punctuations to recognized texts. On the other hand, in several tasks based on automatic speech transcription – such as close captioning / automatic subtitling (Varga et al., 2015), voice mining in audio archives, etc. – missing punctuation marks make reading and interpretation difficult as they require an increased mental effort. Text analysis tools (POS taggers, dependency parsers, etc.) also highly rely on punctuation marks, which may be missing from texts obtained via speech recognition.

Basically two approaches exist in punctuation recovery: (i) prosody based (Christensen et al., 2001) and (ii) language modelling (Batista et al., 2008; Gravano et al., 2009) or recently, sequence modelling (Tilk & Alumäe, 2015) based approaches. The two can be combined into hybrids, as well. Prosody based approaches work fast, whereas language modelling based approaches are usually more resource demanding, which makes their application difficult in online (real-time) automatic speech recognition tasks, which already require much computation for the speech recognition. Moreover, language modelling based approaches show higher language dependency and may be used less easily for spontaneous speech, as they suffer from problems caused by ungrammatical words or non-verbal feedback expressions (Markó–Gósy–Neuberger, 2014) so characteristic in spontaneous speech.

Our interest is to explore punctuation recovery for Hungarian based on speech prosody, keeping in mind the above mentioned considerations, i. e. we would like to propose a fast and efficient approach in terms of both computational requirements and adaptability for spontaneous speech. Prosody based approaches typically focus on some prosodic markers related to punctuations, i.e. look for acoustic markers in duration, fundamental frequency or energy (c.f. Christensen et al., 2001). Unlike most of the studies in the field, which use direct acoustic markers, we intend to incorporate an abstract level of prosodic modelling beside using the acoustic features which relate directly to speech prosody. We do this lead by the conviction that prosody should be considered not only at the layer of different acoustic markers, which relate to different events (i.e a stress or a word boundary), but also as a coherent structure imposed onto the utterance. In Vicsi-Szaszák (2010), a framework was proposed to recover intonational and even phonological phrases directly from speech. This approach relies on modelling phonological phrases with a Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) hybrid, and uses a Viterbi alignment to match the most likely phonological phrase sequence against a speech utterance, using not more than 7 different models (including silence) and fundamental frequency and energy as input acoustic features. This approach reaches ~90% precision and recall in Hungarian phonological phrase detection task with a time accuracy within a length of a syllable (Szaszák et al., 2016).

In our contribution, we prove the hypothesis that the phonological phrase sequence shows characteristic patterns for different punctuation marks, especially regarding within sentence (mostly commas, but also semicolons or dashes) and sentence terminal (period, question or exclamation mark) punctuation marks. The next step is an attempt to model these sequential characteristics with Recurrent Neural Networks (RNN) using Long-Short Term Memory (LSTM) cells, and predict the probability of punctuation marks in a sequence labelling approach. By the implementation of the RNN, we keep a simple structure in order to allow for fast operation. For read speech, this approach is expected to yield the punctuation marks (or probability scores for these punctuation marks) required within a sentence, whereas in spontaneous speech, the approach is a candidate for automatically detecting virtual sentences (Gósy, 2008), and also to identify modality (declarative or interrogative).

We intend to evaluate the RNN in terms of punctuation recovery (precision and recall) both in read speech tasks and in spontaneous speech task. In the latter, a subjective evaluation test is required to analyse the proposed „segmentation" for virtual sentences.

References

Batista, F., Caseiro, D., Mamede, N., & Trancoso, I. (2008). Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news. *Speech Communication*, 50(10), 847–862.

Christensen, Heidi, Yoshihiko Gotoh, & Steve Renals (2001). "*Punctuation annotation using statistical prosody models*." ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding.

Gósy, M. & Kovács, M. 2008. Virtual Sentences of Spontaneous Speech: Boundary Effects of Syntactic-Semantic-Prosodic Properties. In *Human Factors and Voice Interactive Systems*. Springer US, 361–379.

Gravano, A., Jansche, M., & Bacchiani, M. (2009). Restoring punctuation and capitalization in transcribed speech. In *Acoustics, Speech and Signal Processing, 2009*. ICASSP 2009. IEEE International Conference on, 4741–4744.

Markó, A., Gósy, M. & Neuberger, T. (2014). "Prosody patterns of feedback expressions in Hungarian spontaneous speech." *Speech Prosody 2014*, Dublin.

Szaszák, G., Tündik, M.Á., Gerazov, B., Gjoreski, A. (2016). 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.

Tilk, O., & Alumäe, T. (2015). LSTM for punctuation restoration in speech transcripts. In *Interspeech*, 683–687.

Varga, A., Tarján, B., Tobler, Z., Szaszák, G., Fegyó, T, Bordás, C., Mihajlik, P. (2015) Automatic Close Captioning for Live Hungarian Television Broadcast Speech: A Fast and Resource-Efficient Approach. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE 9319. 105–112.

Vicsi, K., Szaszák, G. (2010) Using prosody to improve automatic speech recognition. *Speech Communication* 52:(5) pp. 413–426.

Vicsi, K., Velkei, Sz., Szaszák, G., Borostyán, G., Gordos, G. (2006) Folyamatos, középszótáras beszédfelismerő rendszer fejlesztési tapasztalatai: kórházi leletező beszédfelismerő. *Hiradástechnika* 61:(3) 14–21.