Exploiting prosodic and word embedding based features for automatic summarization of highly spontaneous Hungarian speech

György Szaszák[1] and András Beke[2]
[1] Dept. of Telecommunication and Media Informatics,
Budapest University of Technology and Economics, Hungary
[2] Research Institute for Linguistics, HAS, Hungary

In this contribution, the authors address speech summarization for Hungarian, using highly spontaneous speech material. As the first step, the audio signal is transcribed using an Automatic Speech Recognizer, and speech summarization is carried out on these transcriptions using various text analysis methods. From the speech stream, we also exploit prosody based information, which is used for tokenization prior to textual analysis. We evaluate this prosody based tokenization approach against human performance. The so obtained intonational phrase like tokens are then converted into virtual sentences, and analyzed by the syntactic parser to help ranking based on thematic terms and sentence position. The thematic term is expressed in two ways: TF-IDF and Latent Semantic Indexing. Word embeddings can also be exploited for more robust thematic term calculation. The sentence scores are calculated as a linear combination of the thematic term score and a positional score. The final summary is generated from the top N candidates. Results show that prosody based tokenization reaches human average performance. Audio summarization shows 0.62 recall and 0.79 precision by an F-measure of 0.68, compared to human reference (N=10). Taking into account the high spontaneity of the speech, this results are very encouraging. A subjective test is also carried out on a Likert-scale to allow for a more complete evaluation.

Speech can be processed automatically in several application domains, including speech recognition, speech-to-speech translation, speech synthesis, spoken term detection, speech summarization etc. These application areas use successfully automatic methods to extract or transform the information carried by the speech signal. However, the most often formal, or at least standard speaking styles are supported and required by these applications. The treatment of spontaneous speech (Neuberger et al., 2014) constitutes a big challenge in spoken language technology, because it violates standards and assumptions valid for formal speaking style or written language and hence constitutes a much more complex challenge in terms of modelling and processing algorithms.

Automatic summarization is used to extract the most relevant information from various sources: text or speech. Speech is often transcribed and summarization is carried out on text, but the automatically transcribed text contains several linguistically incorrect words or structures resulting both from the spontaneity of speech and/or speech recognition errors. To sum up, spontaneous speech is "ill-formed" and very different from written text: it is characterized by disfluencies, filled pauses, repetitions, repairs and fragmented words, but behind this variable acoustic property, syntax can also deviate from standard.

Another challenge originates in the automatic speech recognition step. Speech recognition errors propagate further into the text-based analysis phase. Whereas word error rates in spoken dictation can be as low as some percents, the recognition of spontaneous speech is a hard task due to the extreme variable acoustics (including environmental noise, especially overlapping speech) and poor coverage by the language model and resulting high perplexities (Szarvas et al., 2000). To overcome these difficulties, often lattices or confusion networks are used instead of 1-best ASR hypotheses (Hakkani-Tür et al, 2006).

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 (Liu–Xie, 2008). The difficulty comes not only from incomplete structure (often identifying a sentence is already problematic) and recognition errors, but also from missing punctuation marks, which would be fundamental for syntactic parsing and POS-tagging. Speech prosody is known to help in speech segmentation and speaker or topic segmentation tasks (Shriberg et al., 2000). In current work we propose and evaluate a prosody based automatic tokenizer which recovers intonational phrases (IP) and use these as sentence like units in further analysis. Summarization will also be compared to a baseline version using tokens available from human annotation. The baseline tokenization relies on acoustic (silence) and syntactic-semantic interpretation by the human annotators.

In addition, a word-embedding based approach is also considered for speech summarization. Word embeddings project individual words into a semantic space, where words with similar meaning are grouped together. Moreover, such semantic spaces are also able to represent inherent logic linked to meaning and can be used for analogical reasoning tasks or representations (Mikolov et al., 2013). We exploit word embeddings for grouping words with similar meaning in the semantic space, and introduce this knowledge into the thematic term calculation process.

Other research showed that using speech-related features beside textual-based features can improve the performance of summarization (Maskey–Hirschberg, 2005). Prosodic features such as speaking rate; minimuma, maximuma, mean, and slope of fundamental frequency and those of energy and utterance duration can also be exploited. Some approaches prepare the summary directly from speech, relying on speech samples taken from the spoken document (Maskey–Hirschberg, 2006).

References

Hakkani-Tür, D., Bechet, F., Riccardi, G., and Tür, G. (2006). Beyond asr 1-best: using word confusion networks in spoken language understanding. *Computer Speech and Language*, 20(4):495–514.

Liu, Y. and Xie, S. (2008). Impact of automatic sentence segmentation on meeting summarization. In *Proc. Acoustics, Speech and Signal Processing*, ICASSP 2008. IEEE International Conference on, 5009–5012.

Maskey, S. and Hirschberg, J. (2005). Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *INTERSPEECH*, 621–624.

Maskey, S. and Hirschberg, J. (2006). Summarizing speech without text using hidden Markov models. In *Proceedings of the Human Language Technology Conference of the NAACL*, Companion Volume: Short Papers, 89–92.

Neuberger, T., Gyarmathy, D., Gráczi, T. E., Horváth, V., Gósy, M., and Beke, A. (2014). Development of a large spontaneous speech database of agglutinative Hungarian language. In Text, *Speech and Dialogue*, 424–431.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1):127–154.

Szarvas, M., Fegyó, T., Mihajlik, P., and Tatai, P. (2000). Automatic recognition of Hungarian: Theory and practice. Int. *Journal of Speech Technology*, 3(3):237–251.