Summarization of Spontaneous Speech using Automatic Speech Recognition and a Speech Prosody based Tokenizer

György Szaszák[1], Máté Ákos Tündik[1] and András Beke[2]

[1]*Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics,*
2 Magyar tudósok krt., 1117 Budapest, Hungary

[2]*Dept. of Phonetics, Research Institute for Linguistics of the Hungarian Academy of Sciences,*
33 Benczúr utca, 1068 Budapest, Hungary

Keywords: Audio, Speech, Summarization, Tokenization, Speech Recognition, Latent Semantic Indexing.

Abstract: This paper addresses speech summarization of highly spontaneous speech. The audio signal is transcribed using an Automatic Speech Recognizer, which operates at relatively high word error rates due to the complexity of the recognition task and high spontaneity of speech. An analysis is carried out to assess the propagation of speech recognition errors into syntactic parsing. We also propose an automatic, speech prosody based audio tokenization approach and compare it to human performance. The so obtained sentence-like tokens are analysed 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. The sentence scores are calculated as a linear combination of the thematic term score and a positional score. The summary is generated from the top 10 candidates. Results show that prosody based tokenization reaches human average performance and that speech recognition errors propagate moderately into syntactic parsing (POS tagging and dependency parsing). Nouns prove to be quite error resistant. Audio summarization shows 0.62 recall and 0.79 precision by an F-measure of 0.68, compared to human reference. A subjective test is also carried out on a Likert-scale. All results apply to spontaneous Hungarian.

1 INTRODUCTION

Speech is a vocalized form of the language which is the most natural and effective method of communication between human beings. 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 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). In the current paper we are also interested in the as-

Szaszák, G., Tündik, M. and Beke, A.
Summarization of Spontaneous Speech using Automatic Speech Recognition and a Speech Prosody based Tokenizer.
DOI: 10.5220/0006044802210227
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 221-227
ISBN: 978-989-758-203-5

sessment of speech recognition error propagation into text based syntactic parsing – POS-tagging and dependency analysis. Since POS-tagging and especially nouns play an important role in summarization, the primary interest is to see how these are affected by speech recognition errors.

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 and 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 the easier approach, speech summarization is made equivalent to summarizing text transcripts of speech, i.e., no speech-related features are used after the speech-to-text conversion took place. However, the transcribed speech can be used to summarize the text (Christensen et al., 2004). Textual features applied for analysis include the position of the sentence within the whole text, similarity to the overall story, the number of named entities, semantic similarity between nouns estimated by using WordNets (Gurevych and Strube, 2004) etc.

Other research showed that using speech-related features beside textual-based features can improve the performance of summarization (Maskey and 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 and Hirschberg, 2006).

This paper is organized as follows: first the automatic speech recognizer is presented with analysis of error propagation into subsequent processing phases. Thereafter the prosody based tokenization approach is presented. Following sections describe the summarization approach. Finally, results are presented and discussed, and conclusions are drawn.

2 AUTOMATIC SPEECH RECOGNITION AND ERROR PROPAGATION

Traditional Automatic Speech Recognition (ASR) systems only deal with speech-to-text transformation, however, it becomes more important to analyse the speech content and extract the meaning of the given utterance. The related research field is called Spoken Language Understanding. Nowadays, an increasing proportion of the available data are audio recordings, so it becomes more emphasized to find proper solutions in this field. There are two possibilities to analyse the speech content; information can be recovered directly from the speech signal or using parsers after the speech-to-text conversion (automatic speech recognition). This latter option also entails the potential to use analytic tools, parsers available for written language. However, as mentioned in the introduction, a crucial problem can be error propagation – of word insertions, deletions and confusions – from the speech recognition phase.

Our first interest is to get a picture about ASR error propagation into subsequent analysis steps. We select 535 sentences randomly from a Hungarian television broadcast news speech corpus, and perform ASR with a real-time close captioning system described in (Tarján et al., 2014). ASR in our case yields close to 35% Word Error Rate (WER). Please note that the system presented in (Tarján et al., 2014) is much more accurate (best WER=10.5%) , we deliberately choose an operating point with higher WER to approximate a behaviour for spontaneous speech. Punctuation marks are restored manually for this experiment.

Part-Of-Speech (POS) tagging and syntactic parsing is then performed with the *magyarlánc* toolkit (Zsibrita et al., 2013) both on ASR output and the reference text (containing the correct transcriptions). Both outputs are transformed into a separate vector space model (bag of words). Error propagation is estimated based on similarity measures between the two models, with or without respect to word order. In the latter model, the *dogs hate cats* and *cats hate dogs* short documents would behave identically. Although meaning is obviously different, both forms are relevant for queries containing dog or cat. In our case, not word, but POS and dependency tags are compared, by forming uni- and bi-grams of them sequentially.

We use cosine similarity to compare the parses (for N dimensional vectors a and b):

$$sim_{cos}(a,b) = \frac{\sum\limits_{i=1}^{N} a_i b_i}{\sqrt{\sum\limits_{i=1}^{N} a_i^2}\sqrt{\sum\limits_{i=1}^{N} b_i^2}} \qquad (1)$$

The cosine similarity of POS-unigrams and POS-bigrams was found 0.90 and 0.77 respectively for unigrams and bigrams formed from bag-of-words for POS tags of the ASR output and true transcripton, which are quite high despite the ASR errors. Likewise, the result for the dependency labels shows high correspondence between parses on ASR output and true transcription; cosine similarity is 0.89, the standard Labeled Attachment Score and Unlabeled Attachment Score metrics (Green, 2011) (these are defined only for sentences with same length) gave 80% and 87% accuracy. Moreover, the root of the sentence, which is usually the verb, hits 90% accuracy.

Taking the true transcription as reference we also calculate POS-tag error rates analogously to word error rate (instead of words we regard POS-tags). POS-tag error rate is found to be 22% by 35% WER. Checking this selectively for nouns only, POS-tag error rate is even lower: 12%.

These results tell us good news about ASR error propagation into subsequent analysis steps: POS-tags and especially nouns are less sensitive to ASR errors. Checking further for correlation between WER and similarity between parses of ASR output and true transcription, we can confirm that this is close to linear (see Figure 1). This means that by higher WER of the ASR, similarity between true transcription and ASR output based parses can be expected to degrade linearly proportional to WER (see (Tündik and Szaszák, 2016) for a similar experiment in Hungarian), i.e. there is no breakdown point in WER beyond which parsing performance would drop more drastically than does the WER.

2.1 ASR for Summarization

For the summarization experiments, we use different ASR acoustic models (tuned for spontaneous speech) trained on 160 interviews from BEA (Hungarian language), accounting for 120 hours of speech (the interviewer discarded) with the Kaldi toolkit. 3 hidden layer DNN acoustic models are trained with 2048 neurons per layer and tanh non-linearity. Input data is 9x spliced MFCC13 + CMVN +LDA/MLLT. A trigram language model is also trained on transcripts of the 160 interviews after text normalization, with Kneser-Ney smoothing. Dictionaries are obtained using a rule-based phonetizer (spoken Hungarian is very close to the written form and hence, a rule based phonetizer is available).

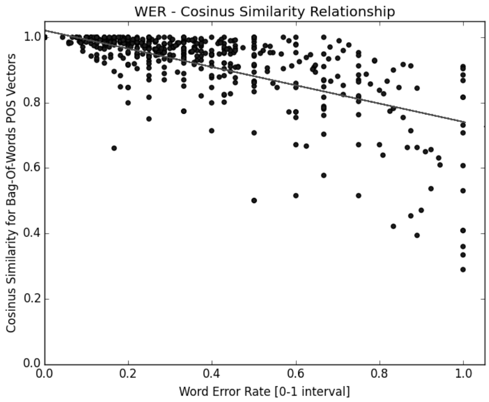

Figure 1: Cosine similarity for POS-tags between true and ASR transcriptions depending on WER. A linear regression line is also plotted.

Word Error Rate (WER) was found around 44% for the spontaneous ASR task. This relative high WER is justified by the high spontaneity of speech.

3 SPEECH PROSODY BASED TOKENIZATION

A speech segmentation tool which recovers automatically phonological phrases (PP) was presented in (Szaszák and Beke, 2012). A PP is defined as a prosodic unit of speech, characterized by a single stress and corresponds often to a group of words belonging syntactically together. The speech segmentation system is based on Hidden Markov Models (HMM), which model each possible type (7 altogether) of PPs based on input features such as fundamental frequency of speech (F0) and speech signal energy. In segmentation mode, a Viterbi alignment is performed to recover the most likely underlying PP structure. The PP alignment is conceived in such a manner that it encodes upper level intonational phrase (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, but giving a perceptually equivalent impression of a short silence in human listener). The algorithm is described in detail in (Szaszák and Beke, 2012), in this work we use it to obtain sentence-like tokens which are then fed into the ASR. We use this IP tokenizer in an operating point with high precision (96% on read speech) and lower recall (80% on read speech) as we consider less problematic missing a token boundary (merge 2 sentences) than inserting false ones (splitting the sentence into 2 parts).

The performance of prosody based tokenization is compared to tokenization obtained from human annotation. Results are presented in section 5.

4 THE SUMMARIZATION APPROACH

After the tokenization for sentence-like intonational phrase units and automatic speech recognition took place, text summarization is split into three main modules. The first module preprocesses the output of the ASR, the second module is responsible for sentence ranking, and the final module generates the summary. This summarization approach is based on (Sarkar, 2012), but we modify the thematic term calculation method. The overall scheme of the system is depicted in Fig. 2.

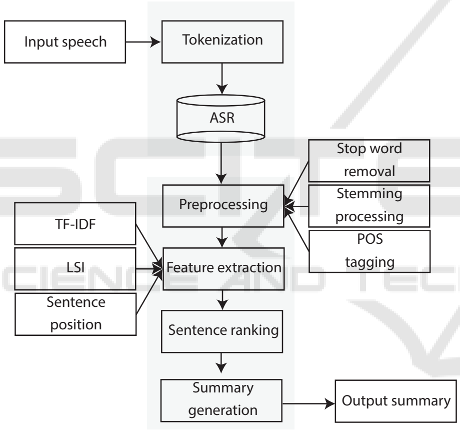

Figure 2: Block scheme of the speech summarization system.

4.1 Pre-processing

Stop words are removed from the tokens and stemming is performed. Stop-words are collected into a list, which contains (i) all words tagged as fillers by the ASR (speaker noise) and (ii) a predefined set of non-content words such as articles, conjunctions etc.

Hungarian is a highly agglutinating language, with a very rich morphology, and consequently, grammatical relations are expressed less by the word order but rather by case endings (suffixes). The *magyarlánc* toolkit (Zsibrita et al., 2013) was used for the stemming and POS-tagging of the Hungarian text. Stemming is very important and often ambiguous due to the mentioned rich morphology. Thereafter, the POS-tagger module was applied to determine a word as corresponding to a part-of-speech. The words are filtered to keep only nouns, which are considered to be the most relevant in summarization.

4.2 Textual Feature Extraction

In order to rank sentences based on their importance, some textual features are extracted:

4.2.1 TF-IDF

(Term Frequency - Inverse Document Frequency) reflects the importance of a sentence and is generally measured by the number of keywords present in it. TF-IDF is a useful heuristic for ranking the words according to their importance in the text. The importance value of a sentence is computed as the sum of TF-IDF values of its constituent words (in this work: nouns) divided by the sum of all TF-IDF values found in the text. TF shows how frequently a term occurs in a document divided by the length of the document, whereas IDF shows how important a term is. In raw word frequency each word is equally important, but, of course, not all equally frequent words are equally meaningful. For this reason it can be calculated using the following equation:

$$IDF(t) = ln\frac{C(all \quad documents)}{C(documents \quad containing \quad term \quad t)} \tag{2}$$

where $C(.)$ is the counting operator. TF-IDF weighting is the combination of the definitions of term frequency and inverse document frequency, to produce a composite weight for each term in each document, calculated as a dot product:

$$TF\text{-}IDF = TF * IDF. \tag{3}$$

4.2.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) exploits context to try to identify words which have similar meaning. LSA is able to reflect both word and sentence importance. Singular Value Decomposition (SVD) is used to assess semantic similarity.

LSA based summarization needs the calculation of the following data (Landauer et al., 1998):

- Represent the input data in the form of a matrix, where columns contain sentences and rows contain words. In each cell, a measure reflecting the importance of the given word in the given sentence is stored. This matrix is often called input matrix (A).
- Use SVD to capture relationships among words and sentences. In order to do this, the input matrix

is decomposed into 3 constituents (sub-matrices):

$$A = U\Sigma VT, \tag{4}$$

where A is the input matrix U represents the description of the original rows of the input matrix as a vector of extracted concepts, Σ is a diagonal matrix containing scaling values sorted in descending order, and V represents the description of the original columns of input matrix as a vector of the extracted concepts (Landauer et al., 1998).

- The final step is the sentence selection for the summary.

4.2.3 Positional Value

The a priori assumption that the more meaningful sentences can be found at the beginning of the document is generally true. This is even more the case for spontaneous narratives, which are the target of our summarization experiments, as the interviewer usually asks the participants to tell something about their life, job, hobbies. People tend to answer with keywords, then they go into details with those keywords recalled. For the calculation of the positional value, the following equitation was used (Sarkar, 2012):

$$P_k = 1/\sqrt{k} \tag{5}$$

where the P_k is the positional score of k^{th} sentence.

4.2.4 Sentence Length

Sentences are of different length (they contain more or less words) in documents. Usually a short sentence is less informative than a longer one and hence, readers or listeners are more prone to select a longer sentence than a short one when asked to find good summarizing sentences in documents (Sarkar, 2012). However, a too long sentence may contain redundant information. The idea is then to eliminate or de-weight sentences which are too short or too long (compared to an average). If a sentence is too short or too long, it is assigned a ranking score of 0.

4.2.5 Sentence Ranking

The ranking score RS_K is calculated as the linear combination of the so-called thematic term based score S_k and positional score P_k. The final score of a sentence k is:

$$RS_k = \begin{cases} \alpha S_k + \beta P_k, & if L_k \geq L_L \quad \& \quad L_k \leq L_U \\ 0 \quad otherwise, \end{cases} \tag{6}$$

where α is the lower, β is the upper cut-off for the sentence position ($0 \leq \alpha, \beta \leq 1$) and L_L is the lower and L_U is the upper cut-off on the sentence length L_k (Sarkar, 2012).

4.3 Summary Generation

The last step is to generate the summary itself. In this process, the N-top ranked sentences are selected from the text. In current work $N = 10$, so the final text summary contains the top 10 sentences.

5 EXPERIMENTS

For the summarization experiments, we use 4 interviews from the BEA Hungarian Spontaneous Speech database (Neuberger et al., 2014), all excluded form the ASR's acoustic and language model training. Participants talk about their jobs, family, and hobbies. Three of the speakers are male and one of them is female. All speakers are native Hungarian, living in Budapest (aged between 30 and 60). The total material is 28 minutes long (average duration was 7 minutes per participant) and is highly spontaneous.

5.1 Metrics

The most commonly used information retrieval evaluation metrics are precision (PRC) and recall (RCL), which are appropriate to measure summarization performance as well (Nenkova, 2006). Beside recall and precision, we use the F_1-measure:

$$F_1 = \frac{2 * PRC * RCL}{PRC + RCL} \tag{7}$$

The challenge of evaluation consists rather in choosing or obtaining a reference summary. For this research we decided to obtain a set of human made summaries, whereby 10 participants were asked to select up to 10 sentences that they find to be the most informative for a given document (presented also in spoken and in written form). Participants used 6.8 sentences on average for their summaries. For each narrative, a set of reference sentences was created: sentences chosen by at least 1/3 of the participants were added to the reference summary. Overlap among human preferred sentences was ranging from 10% to 100%, with an average overlap of 28%. Sentences are appended to the summaries in the order of their appearance in the original text. When using this reference set for the evaluation of the automatic summaries, we filter stop words, fillers (ASR output) from the latter and require at least a 2/3 overlap ratio for the content words. We will refer to this evaluation approach as *soft comparison*.

An automatic evaluation tool is also considered to obtain more objective measures. The most commonly used automatic evaluation method is ROUGE (Lin, 2004). However, ROUGE performs strict string comparison and hence recall and precision are commonly lower with this approach (Nenkova, 2006). We will refer to this evaluation approach as *hard comparison*.

5.2 Results

Text summarization was then run with 3 setups regarding pre-processing (how the text was obtained and tokenized):

- OT-H: Use the original transcribed text as segmented by the human annotators into sentence-like units.
- ASR-H: Use the human annotated tokens and ASR to obtain the text.
- ASR-IP: Tokenize the input based on IP boundary detection from speech and use ASR transcriptions.

Summary generation is tested for all the 3 setups with both TF-IDF and LSA approaches to calculate the thematic term S_k in Equation (6). Results are shown in Table 1 for soft comparison, and Table 2 for hard comparison.

Table 1: Recall (RCL), precision (PRC) and F_1 – soft comparison.

Soft comparison				
Setup	Method	RCL [%]	PRC [%]	F_1
OT-H	TF-IDF	0.51	0.76	0.61
	LSA	0.36	0.71	0.46
ASR-H	TF-IDF	0.51	0.80	0.61
	LSA	0.49	0.77	0.56
ASR-IP	TF-IDF	0.62	0.79	0.68
	LSA	0.59	0.78	0.65

Table 2: Recall (RCL), precision (PRC) and F_1 – hard comparison.

Hard comparison (ROUGE)				
Setup	Method	RCL [%]	PRC [%]	F_1
OT-H	TF-IDF	0.36	0.28	0.32
	LSA	0.36	0.30	0.32
ASR-H	TF-IDF	0.34	0.29	0.31
	LSA	0.39	0.27	0.32
ASR-IP	TF-IDF	0.33	0.28	0.30
	LSA	0.33	0.32	0.32

Overall results are in accordance with published values for similar tasks (Campr and Ježek, 2015). When switching to the ASR transcription, there is no significant difference in performance regarding the soft comparison, but we notice a decrease (rel. 8%) in the hard one (comparing OT-H and ASR-IP approaches). This is due to ASR errors, however, keeping in mind the high WER for spontaneous speech, this decrease is rather modest. Indeed, it seems that content words and stems are less vulnerable to ASR errors, which is in accordance with our findings presented in Section 2.

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). We believe that these good results with IP tokenization are obtained thanks to the better and more infrmation driven structurization of speech parts when relying on prosody (acoustics).

5.3 Subjective Assessment of Summaries

In a separate evaluation step, volunteers were asked to evaluate the system generated summaries on a Likert scale. Thereby they got the system generated summary as is and had to rate it according to the question "How well does the system summarize the narrative content in your opinion?". The Likert scale was ranging from 1 to 5: "Poor, Moderate, Acceptable, Good, Excellent". Results of the evaluation are shown in Fig. 3. Mean Opinion score was 3.2. Regarding redundancy ("How redundant is the summary in your opinion?") MOS value was found to be 2.8.

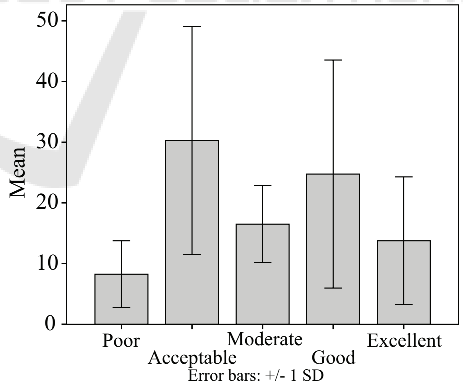

Figure 3: Likert scale distribution of human judgements.

6 CONCLUSIONS

This paper addressed speech summarization for highly spontaneous Hungarian. The ASR transcription and tokenization are sensitive and not yet solved steps in audio summarization. Therefore the authors consider that the proposed IP detection based tokenization is an important contribution, especially as it proved to be as successful as the available human one when embedded into speech summarization. Another basic contribution comes from the estimation of the ASR transcription error propagation into subsequent text processing, at least in terms of evaluating the similarity of POS and dependency tag sequences between human and ASR made transcriptions. Results showed that POS tags and selectively nouns are less sensitive to ASR errors (POS tag error rate was 2/3 of WER, whereas nouns get confused by another part-of-speech even less frequently). Given the high degree of spontaneity of the speech and also the heavy agglutinating property of Hungarian, we believe the obtained results are promising as they are comparable to results published for other languages. 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.

ACKNOWLEDGEMENTS

The authors would like to thank the support of the Hungarian National Innovation Office (NKFIH) under contract IDs *PD-112598* and *PD-108762*.

REFERENCES

Campr, M. and Ježek, K. (2015). Comparing semantic models for evaluating automatic document summarization. In *Text, Speech, and Dialogue*, pages 252–260.

Christensen, H., Kolluru, B., Gotoh, Y., and Renals, S. (2004). From text summarisation to style-specific summarisation for broadcast news. In *Advances in Information Retrieval*, pages 223–237. Springer.

Green, N. (2011). Dependency parsing. In *Proceedings of the 20th Annual Conference of Doctoral Students: Part I - Mathematics and Computer Sciences*, pages 137–142.

Gurevych, I. and Strube, M. (2004). Semantic similarity applied to spoken dialogue summarization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 764.

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.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proc. of the ACL-04 workshop*, volume 8.

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*, pages 5009–5012.

Maskey, S. and Hirschberg, J. (2005). Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *INTERSPEECH*, pages 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*, pages 89–92.

Nenkova, A. (2006). Summarization evaluation for text and speech: issues and approaches. In *INTERSPEECH*, pages 1527–1530.

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*, pages 424–431.

Sarkar, K. (2012). Bengali text summarization by sentence extraction. In *Proc. of International Conference on Business and Information Management ICBIM12*, pages 233–245.

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.

Szaszák, G. and Beke, A. (2012). Exploiting prosody for automatic syntactic phrase boundary detection in speech. *Journal of Language Modeling*, 0(1):143–172.

Tarján, B., Fegyó, T., and Mihajlik, P. (2014). A bilingual study on the prediction of morph-based improvement. In *Proceedings of the 4th International Workshop on Spoken Languages Technologies for Under-Resourced Languages*, pages 131–138.

Tündik, M. A. and Szaszák, G. (2016). Szövegalapú nyelvi elemzö kiértékelése gépi beszédfelismerö hibákkal terhelt kimenetén. In *Proc. 12th Hungarian Conference on Computational Linguistics (MSZNY)*, pages 111–120.

Zsibrita, J., Vincze, V., and Farkas, R. (2013). magyarlanc: A toolkit for morphological and dependency parsing of hungarian. In *Proceedings of RANLP*, pages 763–771.