SPOKEN DOCUMENT SUMMARIZATION USING ACOUSTIC, PROSODIC AND SEMANTIC INFORMATION

Chien-Lin Huang, Chia-Hsin Hsieh and Chung-Hsien Wu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.
{chicco, ngsnail, chwu}@csie.ncku.edu.tw

ABSTRACT

This paper presents a spoken document summarization scheme using acoustic, prosodic and semantic information. First, speech recognition confidence is estimated to choose reliable words from the speech transcription. Prosodic information, including pitch and energy, is used for stressed word selection. Latent semantic indexing (LSI) is adopted to identify significant words. Finally, word trigram and semantic dependency is measured to include the syntactic and semantic information for speech summarization. The dynamic programming (DP) algorithm is used to find the best summarization result according to the summarization score estimated from the above five measures. Finally, the summarized result is presented by the concatenation of the summarized speech words. Experimental results indicate that the proposed approach effectively extracts important words and gives a promising speech summary.

1. INTRODUCTION

In the past years, many efforts have been devoted to text summarization [1]. The major task in text summarization is to analyze the context structure or discourse relation between paragraphs and sentences in an article. Therefore, important sentences are extracted to obtain a text summary. On the other hand, speech summarization is a useful approach that compresses speech utterances into concise and meaningful speech sentences. Generally, speech summarization relies on the transcription from a large-vocabulary continuous-speech recognizer (LVCSR). Summarization result is obtained from the transcription according to the analysis of semantic and syntactic information [2]. The major difference between text and speech summarization is that speech summarization needs to deal with the problem due to speech recognition error. Despite of the recognition error in speech summarization, prosody is the particular information conveying stressed words in spoken documents and can be used to improve the summarization performance.

For previous methods using prosody, Koumpis and Renals [3] proposed a summarization approach for voicemail message using prosodic cues and lexicon content. In [4], prosodic features such as pitch, power and pause are applied to analyze the summary units and their dependency relations.

This study focuses on spoken document summarization of TV news broadcasts. Speech recognition confidence is first estimated to choose reliable words from the speech transcription. Second, since the anchor speech is planned, important words will be specially stressed and the prosodic information will be useful for the extraction of stressed words. Third, latent semantic indexing (LSI) is adopted for significant word extraction. Finally, word trigram and semantic dependency are measured to include the syntactic and semantic information for speech summarization. Based on these five scores, the dynamic programming algorithm is used to summarize the concise and meaningful sentences.

2. AUTOMATIC SPEECH SUMMARIZATION

Given an original speech utterance with N words, the corresponding transcription $X = (w_1, w_2, ..., w_N)$ is obtained using an LVCSR. A topic-related corpus is collected and used to extract important keywords related to the spoken documents for summarization. These keywords in an utterance are used to estimate the compression ratio ∂ . A summarized sentence $Y = (w_1, w_2, ..., w_M)$ with $M = N \times \partial \times 100$ words which maximizes the following summarization score is obtained:

$$S(Y) = \sum_{m=1}^{M} \{\lambda_A A(w_m) + \lambda_C C(w_m) + \lambda_R R(w_m) + \lambda_L L(w_m \mid w_{m-2}, w_{m-1}) + \lambda_B B_{SDG}(w_{m-1}, w_m)\} \quad (1)$$

where the prosody score $A(w_m)$ considers pitch and energy of word w_m . $C(w_m)$ denotes the confidence score of word w_m obtained from the LVCSR. $R(w_m)$ denotes the word significance score. $L(w_m \mid w_{m-2}, w_{m-1})$ represents the trigram probability and $B_{SDG}(w_{m-1}, w_m)$ is the word concatenation score which is obtained using the semantic dependency grammar. As shown in Figure 1, the dynamic programming search algorithm is applied to find the summarized result with the highest summarization score.

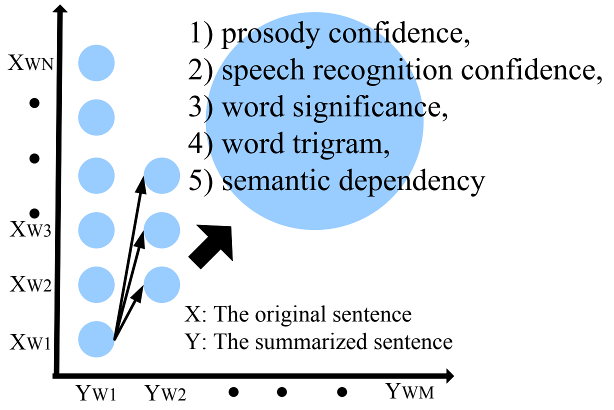

Fig.1: Example of the DP search for summarization

The range of these five scores is normalized to lie between 0 and 1. The weights λ_A, λ_C, λ_R, λ_L and λ_B are used to linearly combine these scores. In the preprocessing step, stop words are disregarded because they are meaningless. Furthermore, the repeated words in the spoken document are discarded because they have the same significance.

2.1. Prosody Confidence

Basically, the prosodic information can reflect a speaker's intent, mood and emotion. This study employs the prosodic information, containing energy and pitch, to extract the stressed words important in the spoken documents.

Energy Measure

In speech communication, higher energy is usually reflected by highly important words for emphasis. In this paper, the energy measure for the word w_m is calculated as follows.

$$e(w_m) = \frac{1}{N_{w_m}} \sum_{i}^{N_{w_m}} \frac{(e_i(w_m) - e_{\min})}{(e_{\max} - e_{\min})} \qquad (2)$$

where N_{w_m} is the total number of frames in the word w_m; $e_i(w_m)$ represents the energy of the i^{th} frame in word w_m; $e_{\min}$ and $e_{\max}$ are the minimum and maximum energies of the spoken document for summarization, respectively.

Pitch Measure

Pitch is commonly used to measure the rate of vocal fold vibration. In speech summarization applications, it is easily observed that stressed words usually correspond to higher pitch. This paper applies an autocorrelation method to estimate the fundamental frequency F_0. Eqs. (3) and (4) define the transfer functions to estimate the pitch measure of the word w_m in a logarithmic scale expressed in semitones [6].

$$p_i = 80\log_{10} F_0(i) \qquad (3)$$

$$p(w_m) = \frac{1}{N_{w_m}} \sum_{i}^{N_{w_m}} 24 * \frac{(p_i(w_m) - p_{\min})}{(p_{\max} - p_{\min})} \qquad (4)$$

where N_{w_m} is the total number of frames in the word w_m; $p_i(w_m)$ represents the pitch of the i^{th} frame in word w_m, and $p_{\min}$ and $p_{\max}$ represent the minimum and maximum pitches in the spoken document, respectively.

A linear combination method is applied to calculate the prosody significance.

$$A(w_m) = \alpha_1 e(w_m) + \alpha_2 p(w_m) \qquad (5)$$

where α_1 and α_2 are the weighting parameters for balancing the energy and pitch measures. Figure 2 shows an example of the prosody measures for the sentence "提前 (in advance) 歡慶 (to celebrate) 父親節 (Father's Day)." In this spoken utterance, "父親節 (Father's Day)" is more important than other words and presents a higher prosody confidence.

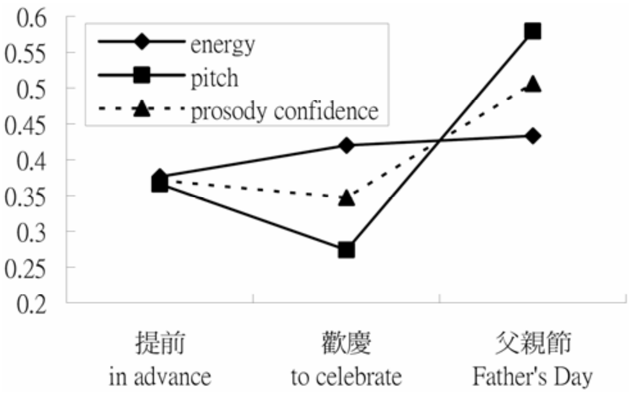

Fig. 2: Example of prosody score

2.2. Speech Recognition Confidence

Speech recognition confidence is defined as the posterior probability of each transcribed word and is widely used to evaluate the assurance of the recognition result. It is calculated using a word graph obtained by a decoder and used as the confidence measure [7].

2.3. Word Significance

Word significance is to evaluate the importance of the words in an utterance. A corpus consisting of the articles and their corresponding title keywords related to the spoken documents for summarization was collected to estimate the word significance. As in [5], a topic-related document retrieval model is applied to obtain the most relevant document and the corresponding title keywords. After retrieving the most relevant document d^*, the words in the corresponding title t^* contain the most important information related to document d^*. The word significance score $R(w_m)$ of w_m in the transcribed sentence is calculated according to the words in title t^* and the word correlation matrix obtained using latent semantic indexing [5]:

$$R(w_m) = \max_b \{P_{LSI}(w_m, w_b^{t^*}) \cdot f_{w_m} \cdot \ln(N / df_{w_m})\} \qquad (6)$$

where $P_{LSI}(w_m, w_b^{t^*})$ denotes the similarity between word w_m and title keyword $w_b^{t^*}$; f_{w_m} is the term frequency of word w_m in the document. df_{w_m} represents the document frequency of word w_m and N denotes the number of documents in the corpus.

2.4. Word Trigram Score

The word trigram score $L(w_m \mid w_{m-2}, w_{m-1})$ is used to estimate the concatenation probability of a word sequence. The trigram probability is interpolated from trigram, bigram and unigram to smooth the frequencies.

2.5. Semantic Dependency Score

The semantic relation between two words is analyzed from the semantic dependency grammar. A modified dependency grammar [8] is introduced to obtain the semantic dependency score $B_{SDG}(w_a, w_b)$ as follows:

$$\begin{aligned} &B_{SDG}(w_a, w_b) \\ &= \frac{1}{N_s} \sum_{j=1}^{N_s} \sum_i \sum_r f_{DR_i^r(w_a, w_b)}(T_i, S^j(w_a, w_b)) \times f_{T_i}(S^j(w_a, w_b)) \end{aligned} \qquad (7)$$

where $S^j(w_a, w_b)$ denotes sentence S^j containing words w_a and w_b. $f_{T_i}(.)$ denotes the probabilistic context free grammar (PCFG). T_i denotes the parse tree. $f_{DR_i^r}(.)$ denotes the score of SDG. N_s denotes total sentence numbers. Dependency graph $D_i = \{DR_i^r(w_a, w_b) \mid 1 \le r \le N_w - 1\}$ represents the set of dependency relations DR_i^r from the parse tree T_i and sentence $S^j(w_a, w_b)$ with N_w words. Figure 3 shows an example of a dependency graph for the sentence "老太太 (The old lady) 很 (greatly) 想 (misses) 台北 (Taipei)." In this example, *Degree* and *Experience* represent two dependency relations between two words.

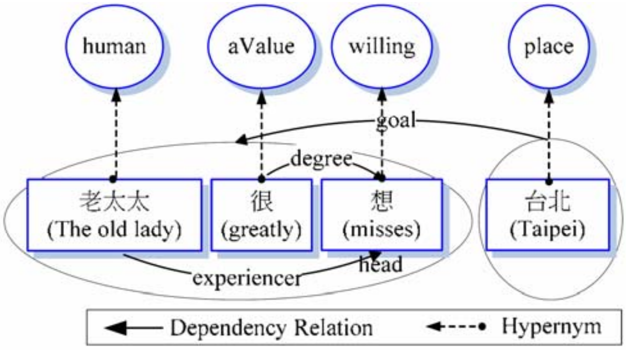

Fig.3: Example of the semantic dependency graph

In order to avoid the sparse data problem when estimating dependency relations, word hypernyms based on HowNet [10] are used in PCFG.

$$f_{DR_i^r(w_a, w_b)}(T_i, S^j(w_a, w_b)) \cong f_{DR_i^r(H(w_a), H(w_b))}(T_i, S^j(w_a, w_b)) \qquad (8)$$

where $H(w_a)$ denotes the hypernym of w_a, Furthermore, the score is estimated using the following equation:

$$\begin{aligned} &f_{DR_i^r(H(w_a), H(w_b))}(T_i, S^j(w_a, w_b)) \\ &= C(DR_i^r(H(w_a), H(w_b))) / C(H(w_a), H(w_b)) \end{aligned} \qquad (9)$$

where $C(DR_i^r, H(w_a), H(w_b))$ denotes the frequency that dependency relation $DR_i^r(H(w_a), H(w_b))$ happens in the training corpus. $C(H(w_a), H(w_b))$ denotes the co-occurrences of $H(w_a)$ and $H(w_b)$ in the training corpus

3. EXPERIMENTS

3.1. Experimental Setup

For evaluation, an HMM-based Mandarin LVCSR was constructed. In the speech recognizer, a 39-dimensional mel-frequency cepstral coefficients (MFCC) and a twopass decoder were used. The training data for the acoustic models consist of 4 hour anchor speech from TV news broadcasts from 2001 to 2002. The language model was trained using a newswire text corpus from News website consisting of 20 million Chinese characters in the same time period. The character recognition accuracy achieved about 80%. Furthermore, the semantic dependency grammar was constructed from the Sinica Treebank [9] with 36953 sentences and HowNet knowledgebase [10]. We extracted 22,025 rules according to the tree structure of Part-of-Speeches (POSs) and their corresponding probabilities from the Treebank were obtained.

The evaluation data for speech summarization was collected from TV news broadcast in 2002. There are 255 spoken documents consisting of 2150 utterances by a female anchor speaker.

3.2. Evaluation of Character Accuracy Compared with Manual Summarization Result

The results from automatic speech summarization were compared with the subjective results from manual summarization. Five graduate students were invited to summarize target references from correct news articles and the character accuracy is estimated using the following measure:

$$P_{accuracy} = (W - I - D - S)/W \qquad (10)$$

where W is the number of characters. I , D and S denote the numbers of insertion, deletion and substitution character errors, respectively.

Table I. Character accuracy for different scores

Scores	Accuracy	Insertion	Deletion	Substitution
A	0.35	0.04	0.27	0.34
C	0.42	0.04	0.29	0.25
R	0.45	0.09	0.27	0.19
L	0.46	0.03	0.3	0.21
B	0.33	0.05	0.3	0.32
C+L+R+B	0.52	0.04	0.3	0.14
A+C+L+R+B	0.56	0.03	0.22	0.19

In this table, A denotes the prosody score; C represents the confidence score; R denotes the word significance score; L denotes the linguistic score, and B denotes the semantic dependency score. The values of the combination parameters λ_A , λ_C , λ_R , λ_L and λ_B were empirically chosen as 0.18, 0.21, 0.22, 0.23 and 0.16, respectively, according to the individual performance for each score. Table I shows the experimental results of and the summarization system using the combined score (A+C+L+R+B) obtains a better performance compared to the systems using individual score. Furthermore, the prosodic confidence can provide an improvement of 4% compared to the approach without prosody (C+L+R+B). This reveals prosody is also important for speech summarization.

4. CONCLUSION

This study has presented an approach for speech summarization using acoustic, prosodic, semantic information for spoken document summarization. Five scores, including speech recognition confidence, prosody score, word significance score, word trigram score and semantic dependency score, are estimated and used for speech summarization. The DP algorithm is used to find the best summarization result. Experimental results demonstrate that the proposed framework achieves a satisfactory performance.

5. REFERENCES

[1] I. Manu and M. Maubury, *Advances in Automatic Summarization*. Cambridge, MA: MIT Press, 1999.

[2] C. Hori and S. Furui, “A new approach to automatic speech summarization,” *IEEE Trans. on Multimedia*, vol. 5, no. 3, pp. 368-378, 2003.

[3] K. Koumpis and S. Renals, “The role of prosody in a voicemail summarization system,” in *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, USA, 2001

[4] Kiyonori Ohtake, Kazuhide Yamamoto, Yuji Toma, Shiro Sado, Shigeru Masuyama and Seiichi Nakagawa, "Newscast Speech Summarization via Sentence Shortening based on Prosodic Features," *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 167-170, 2003

[5] C.H. Hsieh, C.L. Huang and C.H. Wu, “Spoken document summarization using topic-related corpus and semantic dependency grammar,” in *Proc. ISCSLP’04*, Hong Kong, 2004, pp. 333-336.

[6] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, "Spoken Language Processing," Prentice Hall, Inc., 2001

[7] T. Kemp and T. Schaaf, “Estimating confidence using word lattices,” *in Proc. 5th Eurospeech*, vol. 2, Rhodes, Greece, 1997, pp. 827-830.

[8] Christopher D. Manning and Hinrich Schutze, “Foundations of Statistical Natural Language Processing,” The MIT Press, 1999

[9] CKIP Treebank http://godel.iis.sinica.edu.tw/CKIP/treebank/

[10] HowNet, http://www.keenage.com/

[11] M. Banko, V. Mittal and M. Witbrock, “Headline generation based on statistical translation,” *in Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics,* 2000, pp. 318-325.