News Dictation and Article Classification Using Automatically Extracted Announcer Utterance

Yasuo Ariki Jun Ogata Masafumi Nishida

Faculty of Science and Technology, Ryukoku University,
1-5 Yokotani, Oe-cho, Seta, Otsu-shi, 520-2194 Japan,
ariki@rins.ryukoku.ac.jp

Abstract. In order to construct a news database with a function of video on demand (VOD), it is required to classify news articles into topics. In this study, we describe a system which can dictate news speech, extract keywords and classify news articles based on the extracted keywords. We propose that it is sufficient to dictate only the announcer utterance in classifying the news articles and it contributes to reduce the processing time. As an experiment, we compared the classification performance of news articles in two cases; dictating only the announcer utterances which are automatically extracted and dictating a whole speech which includes reporter or interviewer utterances.

1 Introduction

Recently, TV news programs are broadcasted all over the world owing to the broadcast digitization. In this situation, TV viewers require to select and watch the most interesting news. For that purpose, word indexing and article classification are key techniques.

The word indexing is a technique to put the discriminative keywords on the news speech articles in order to retrieve the interesting articles. On the other hand, the article classification is a technique to classify the news speech articles into groups (topics) based on their contents such as politics, economy, science and sports in order to retrieve the same kind of articles[?]. These two techniques are strongly required because manually indexing and classification are almost impossible.

From this viewpoint, we propose in this paper a method to automatically index and classify TV news articles into 10 topics based on a speech dictation technique using speaker independent phoneme HMMs and word bigram. After the dictation of the spoken news articles, pre-defined keywords are searched and the new articles are classified based on the keywords.

In general, news speech includes reporter speech as well as announce speech. The announcer speech is clear but the reporter speech sometimes noisy due to wind or environmental noises so that the dictation accuracy for the reporter speech is lower than for the announcer speech. Therefore if the speech dictation process is applied only to the announcer speech, we can reduce the processing time without decreasing the news classification accuracy.

From this viewpoint, we propose in this paper a method to automatically divide the TV news speech into speaker sections and then index in real time who is speaking. This can be realized by using a technique of speaker verification[**?**]. However, the speaker verification technique is sensitive to the time lapse. Namely, the speech characteristics of each speaker is subject to change day by day. To solve this problem, speaker models are not prepared in advance but are constructed through indexing in self-organization mode.

We verified the effectiveness of our proposed methods by carrying out the experiment in extracting and dictating only the announcer speech and then classifying the news articles into 10 topics.

2 Announcer Speech Extraction

2.1 Speaker Verification

Speaker verification is a technique to judge if the input speech belongs to the specified person or not[**?**]. Fig.1 shows the speaker verification process. When the speaker ID of speaker A and his speech are fed to the verification system, the distance is computed between the model of the speaker A and the input speech. If the distance is smaller than some threshold, the input speaker is accepted as the true speaker A. Otherwise the input speaker is rejected. In our experiment, speaker subspace is constructed as the speaker model and the distance between the input speech and the speaker subspace is computed.

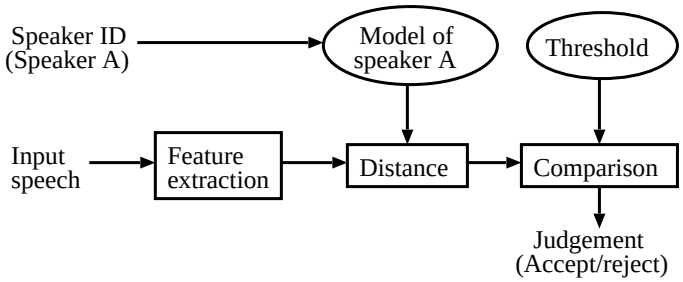

Fig. 1. Speaker verification

2.2 Speaker Subspace

As shown in Fig.**??**, we observe speech data $X^{(i)}$ of the speaker i and speech data $X^{(j)}$ of the speaker j in an observation space. The speech data are a sequence of spectral feature vectors $x_t^{(i)}$ and $x_t^{(j)}$ obtained at time t by short time spectral analysis. We denote the speech data $X^{(i)}$ as a matrix whose row is a spectral

feature vector ${x_t^{(i)}}^T - {\mu^{(i)}}^T$ $(1 \leq t \leq M)$. Here $x_t^{(i)}$ denotes an observed feature vector and $\mu^{(i)}$ is their mean vector. The column of the matrix corresponds to frequency f $(1 \leq f \leq N)$.

By singular value decomposition, the speech data matrix $X^{(i)}$ is decomposed as

$$X^{(i)} = U^{(i)} \Sigma^{(i)} {V^{(i)}}^T \tag{1}$$

Here $U^{(i)}$ and $V^{(i)}$ are the matrices whose columns are eigenvectors of $X^{(i)} {X^{(i)}}^T$ and ${X^{(i)}}^T X^{(i)}$ respectively. $\Sigma^{(i)}$ is the singular value matrix of $X^{(i)}$.

The eigenvectors of the correlation matrix ${X^{(i)}}^T X^{(i)}$ are the orthonormal bases of the speech data $X^{(i)}$, computed based on a criterion that the total distance is minimized between feature vectors $x_t^{(i)} - \mu^{(i)}$ and the orthonormal bases[**?**]. Then $V^{(i)}$ is considered as orthonormal bases of the speaker space. This is completely same as the principal component analysis of the speech data $x_t^{(i)}$ [**?**].

If the large singular values up to r numbers are selected from the matrix $\Sigma^{(i)}$, the matrix $V^{(i)}$ becomes N x r dimension and is considered as the speaker subspace.

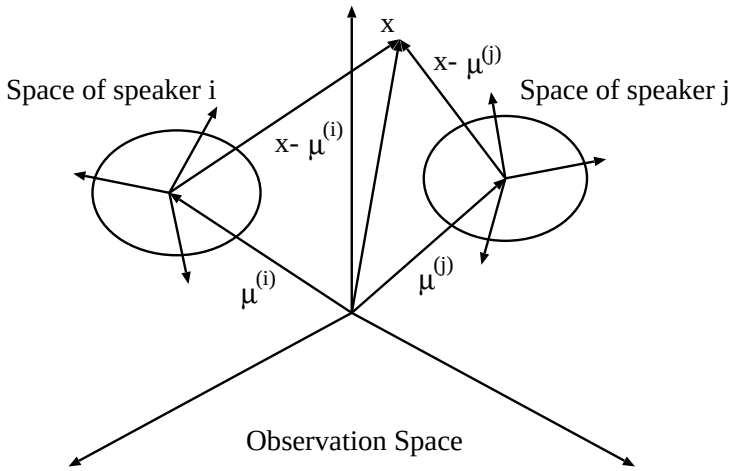

Fig. 2. Speaker subspace

2.3 Verification by Speaker Subspace

The speaker subspace $V^{(i)}$ is composed of orthonormal bases $\{v_1^{(i)}, \cdots, v_r^{(i)}\}$ of the speech data $X^{(i)}$. Speaker verification can be carried out by computing a

distance from an input speech vector x_t in the observation space to the speaker subspace $V^{(i)}$.

The distance is presented as follows using a projection matrix $P^{(i)}$ from the observation space to the speaker subspace.

$$\begin{aligned} Dist(V^{(i)}, x_t) &= \|x_t - \{P^{(i)}(x_t - \mu^{(i)}) + \mu^{(i)}\}\|^2 \\ &= \|(I - P^{(i)})(x_t - \mu^{(i)})\|^2 \end{aligned} \tag{2}$$

where the projection matrix $P^{(i)}$ is defined as:

$$P^{(i)} = \sum_{k=1}^{r} v_k^{(i)} {v_k^{(i)}}^T = V^{(i)} {V^{(i)}}^T \tag{3}$$

Equation(**??**) means that the projection matrix $P^{(i)}$ is obtained using the orthonormal bases of the speech data $X^{(i)}$.

The distances computed by Eq.(**??**) between speech vectors x_t and the speaker subspace $V^{(i)}$ are averaged over time t. The speaker is identified as one with the minimum averaged distance between the speech vectors and the subspace.

2.4 Extraction of Speaker Section

Continuous news speech is divided into sections of respective speaker by using the speaker verification technique. The sections are called here " speaker sections". The continuous news speech is also divided into sections separated by silence. The sections are called " speech sections". The extraction process for speaker sections is as follows;

(1) Averaged power is computed at every 1 second on the input speech. If it is lower than some threshold it is regarded as silence. The speech section between two silences is extracted.
(2) Using the firstly extracted speech section, a speaker subspace is constructed. This speaker subspace corresponds to the model of the speaker A shown in Fig.1. Here the threshold θ to accept or reject the speaker is determined in advance as follows, using μ (mean) and σ (standard deviation) of the distance between speech data in the first speech section and the speaker subspace;

$$\theta = \mu + \frac{\sigma}{3} \tag{4}$$

(3) On the successive speech section, the distance is computed between the input speech and the model. If the distance is lower than the threshold θ, it is judged that the speaker A is still speaking. In this case, the speaker subspace model is updated as well as the threshold θ using all the speech data verified as speaker A.
(4) Otherwise, it is regarded that speaker A has finished his speech and new speaker or previous speaker begins speaking. To judge it, the distance between the input speech section and the previously constructed speaker subspace models are computed. If some speakers have lower distance than threshold θ, then the input speaker is judged to be the speaker with the lowest

distance. Otherwise, the input speaker is regarded as a new speaker and step (2) begins starting.

2.5 Experimental Result

We selected 48 news articles which included reporter speech as well as announcer speech from 45 days NHK news program. Everyday news program usually contains 4 or 5 articles and continues for 5 minutes. For these 48 news articles we carried out the experiment to extract the announcer sections. The dimension of speaker subspace was set to 7 after preliminary experiment. The experimental condition is shown in Table.??.

Table 1. Experimental condition

Speech data	48 NHK news articles
Sampling frequency	12kHz
Frame length	20ms
Frame period	5ms
Window type	Hamming window
Features	LPC Cepstrum(16 orders)
Subspace dimension	7
Threshold θ	$\theta = \mu + \frac{\sigma}{3}$

The extraction of announcer sections was evaluated by the extraction rate and the precision rate defined as follows;

$$Extraction\ rate = \frac{\left\{\begin{array}{l}\text{Number of correctly verified}\\ \text{speech sections as announcer}\end{array}\right\}}{\left\{\begin{array}{l}\text{Number of total speech sec-}\\ \text{tions of the announcer}\end{array}\right\}} \tag{5}$$

$$Precision\ rate = \frac{\left\{\begin{array}{l}\text{Number of correctly verified}\\ \text{speech sections as announcer}\end{array}\right\}}{\left\{\begin{array}{l}\text{Number of verified speech}\\ \text{sections as announcer}\end{array}\right\}} \tag{6}$$

Here announcer is judged to be the speaker who spoke the longest time in 1 day 5 minutes NHK news.

The announcer extraction result is shown in Table??. The extraction rate was 92.6% and the precision rate was 82.9%.

Table 2. Experimental result(%)

Extraction rate	92.6
Precision rate	82.9

3 Speech Dictation

3.1 Experimental Condition

We carried out speech diction for the 48 news articles after the extraction of announcer section. The language model is the word bigram produced from MAINICHI Japanese newspaper of 45 months after morphological analysis. The number of the words in the dictionary is 5,000. The word bigram was back-off smoothed after cutting off at 2 words.

Speaker independent 41 monophone HMMs were constructed. Their structure is 5 states with 3 loops and 8 mixtures for each state. They were trained using 21,782 sentences spoken by 137 Japanese males. These speech data was taken from the database of acoustic society of Japan. The acoustic parameters are 39 MFCCs with 12 Mel cepstrum, log energy and their first and second order derivatives. Cepstrum mean normalization is applied to each sentence to remove the difference of input circumstances. Table**??** shows the experimental conditions for acoustic analysis (AA) and HMM.

In the dictation experiment, we used HTK (HMM Toolkit)[**?**] as the decoder which can perform Viterbi decoding with beam search using the above mentioned language model and the acoustic model.

Table 3. Acoustic Analysis(AA) and HMM

	Sampling frequency	16kHz
	High-pass filter	$1 - 0.97z^{-1}$
A	Feature parameter	MFCC (39th)
A	Frame length	20ms
	Frame shift	5ms
	Window type	Hamming window
H	Number of states	5 states 3 loops
M	Learning method	Concatenated training
M	Type	Left to right continuous HMM
	Number of mixtures	8

3.2 Dictation Result

The dictation was carried out for 48 NHK news articles. They were already divided into two speaker sections by the previously described method; announcer and reporter (interviewer). Table?? shows the property of the 48 news articles in terms of announcer section (Anchor), reporter section (Other) and mixed total section (All). In the reporter section, the ratio of unknown words to the 5,000 dictionary words is high compared with other two sections. This is also reflected in the test-set perplexity which is the measure of task complexity.

Table 4. News articles used for dictation

	Anchor	Other	All
Number of sentences	247	116	363
5K unknown word ratio	13.9%	29.3%	20.6%
Test-set perplexity	153.7	285.2	177.6

The dictation result is shown in table??. In the table, word correct rate and word accuracy are defined as follows;

$$\text{Word correct rate} = \frac{N - S - D}{N} \cdot 100 \tag{7}$$

$$\text{Word accuracy} = \frac{N - S - D - I}{N} \cdot 100 \tag{8}$$

S : The number of substituted words
D : The number of deleted words
I : The number of inserted words
N : Total number of words

The word error rate is defined as ($100 - \textit{word accuracy}$). From the table, it can be seen that the announcer speaks clearly and grammatically in the clear circumstance. On the other hand, the reporter speaks colloquially in the noisy circumstance. The dictation result is used for topic classification in the successive process.

Table 5. Dictation result(%)

	Anchor	Other	All
Word error rate	39.7	79.3	54.6
Word correct rate	66.5	23.5	48.7
Word accuracy	60.3	20.7	45.4

4 Classification of News Articles

4.1 Classification Flow

Fig.?? shows the classification flow of news speech articles by speech dictation technique. Before the article classification, news articles in the news program are automatically separated each other using the algorithm mentioned in [?]. In the flow, there are following two phases;

(1) Speech dictation phase: word sequence and their probabilities $Ps(w)$ are computed by applying the speech dictation technique.
(2) Article classification phase: articles are classified by integrating the keyword probability $Ps(w)$ and the topic contribution probability of each word $P(n|w)$.

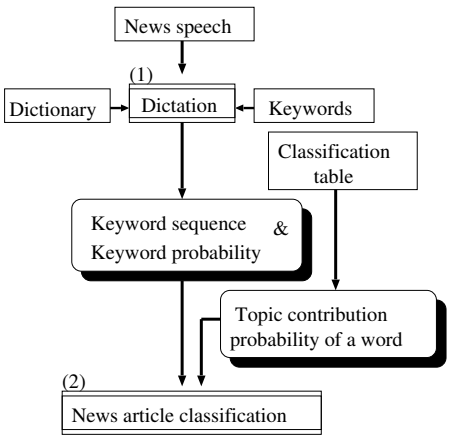

Fig. 3. Flow of news article classification

4.2 Word Probability

Speech dictation by HTK described in ?? can produce the word probability as well as the word sequence for each spoken news sentence. After speech dictation

to the 48 news articles, keywords w_i are searched together with their probability $P_s(w_i)$. The keywords were determined in advance as the words included in both "classification indices for ASAHI newspaper article database" and 48 news articles. The number of keywords was 399 words in this experiment.

4.3 Topic Contribution Probability

In article classification using a keyword sequence, topic contribution of a keyword (TCKW) $P(n|w)$ is computed in advance as shown in Eq.(??). The TCKW indicates how the keyword contributes to identify the topic and is defined as the ratio of the occurrence of the keyword w included in the topic n to the occurrence of the keyword w included in all the topics. This definition is *a posteriori* probability of the topic n conditioned by the word w.

$$P(n|w) = \frac{\left\{\begin{array}{l}\text{The number of occurrence of}\\ \text{the keyword } w \text{ included in the}\\ \text{topic } n\end{array}\right\}}{\left\{\begin{array}{l}\text{The number of occurrence of}\\ \text{the keyword } w \text{ included in all}\\ \text{the topics}\end{array}\right\}} \tag{9}$$

In this study, we used "classification indices for ASAHI newspaper article database" in computing the TCKW. Fig.?? shows a part of the classification indices.

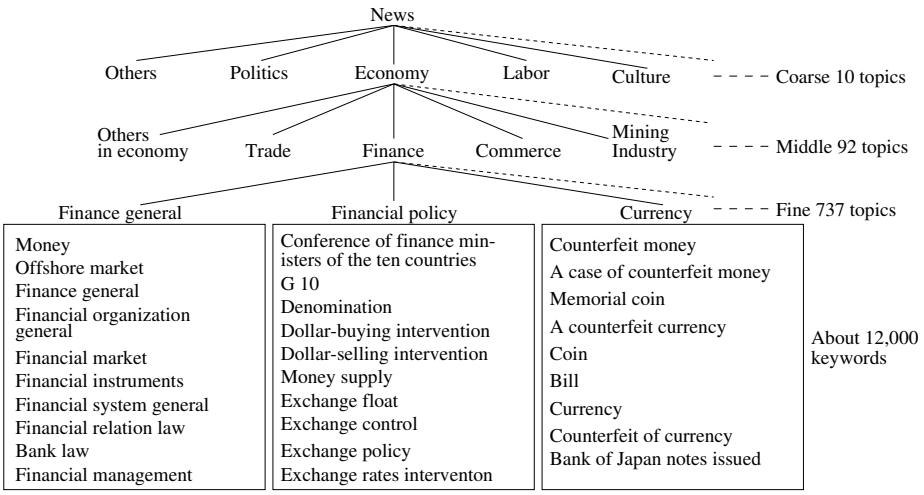

Fig. 4. A part of classification indices

It includes 12,000 keywords and they have links to the related topics. There are three levels in grouping of the topics; coarse, middle and fine level. They have about 10, 92 and 737 kinds of topics respectively. In the fine topics, about 16 indices are prepared at average in each topic. We selected coarse level of 10 topics classification in this study. The 10 topics for classification are Politics, Economy, Labor, Culture, Science, Society, Accidents, Sports, Internationality and Others.

Table?? shows an example of how to compute the TCKW. In the table, there are three complex words which include "Japan-U.S." word in the topic of politics. In the same way, there are two and zero in the topic of economy and society respectively. In total the number of complex words including "Japan-U.S." is five. In this example, the TCKW of the "Japan-U.S." word is computed as follows;

$$P(\text{Politics}|\text{Japan-U.S.}) = \frac{3}{5} = 0.6$$
$$P(\text{Economy}|\text{Japan-U.S.}) = \frac{2}{5} = 0.4$$
$$P(\text{Society}|\text{Japan-U.S.}) = \frac{0}{5} = 0.0$$

The TCKW is computed for all the keywords in advance.

Table 6. Example of topics and keywords

Topic	Japan-U.S.	total
Politics	Japan-U.S. security treaty	3
	Japan-U.S. administrative agreement	
	Japan-U.S. relation	
Economy	Japan-U.S. economic friction	2
	Japan-U.S. trade friction	
Society		0

4.4 Topic Probability

The topic probability $P(n|w_1, \cdots, w_k)$ that the article is classified into the topic n after the extraction of the keywords $w_1, \cdots, w_k$ is shown in Eq.(??).

$$P(n|w_1, \cdots, w_k) = \sum_{i=1,\cdots,k} P(w_i) \times P(n|w_i) \tag{10}$$

where $P(n|w_i)$ is the topic contribution probability of the keyword w_i. The probability $P(w_i)$ is replaced by the normalized word probability as follows;

$$P(w_i) = \frac{Ps(w_i)}{\sum_{j=1,\cdots,k} Ps(w_j)} \tag{11}$$

This topic probability is the integration of acoustic word probability $Ps(w)$ and *a priori* knowledge probability TCKW. The news article can be classified into the topics with the highest topic probability $P(n|w_1,\cdots,w_k)$.

5 Classification Result

Table?? shows the classification result of the 48 news articles. In the table, the classification rate is the ratio of the number of correctly classified articles to the total number of articles. The article is judged to be correctly classified if it is classified into the correct topic. The correct topic is determined by setting the word probability $Ps(w_i) = 1$ for the true keywords which are obtained from the text data.

From the table, it can be seen that the classification rate is 63.6% with 60.3% word accuracy for the announcer speech. On the other hand, the classification rate for all utterances is 63.0% with 45.4% word accuracy. This result indicates that the dictation accuracy for reporter speech is lower than that for the announcer speech due to the noisiness and colloquialism. Even though it is true, the classification rate is almost same. This indicates that the keywords are mainly included in the announcer speech and they are well extracted compared with the reporter speech.

Table 7. Classfication result(%)

	Word accuracy	Classification rate
Anchor	60.3	63.6
All	45.4	63.0

6 Conclusion

We have described the automatic classification system of TV news articles. Keywords were extracted from news speech articles after their dictation using word bigram and speaker independent HMM. The acoustic probabilities of the keywords were multiplied with the topic contribution probabilities which were computed from "classification indices for ASAHI newspaper article database" and

the topic probability of the article was produced. The news speech articles were classified based on this topic probability.

In order to speed up the processing time, we have omitted the reporter speech section and still kept the same classification accuracy. The highest classification rate was 63.6% and seems to be low in accuracy for real application. We need to improve speech dictation technique in future.

Acknowledgment

This research was partly supported by "Research for the Future" Program of Japan Society for the Promotion of Science under the Project "Advanced Multimedia Content Processing" (Project No. JSPS-RFTF97P00501) and partly supported by the Japanese Ministry of Education Grant-in-Aid for Scientific Research on Priority Area: "Advanced databases," area no. 275(08244103).

References

1. Y.Ariki, M.Sakurai and Y.Sugiyama : " Article Extraction and Classification of TV News Using Image and Speech Processing", CODAS96 (International Symposium on Cooperative Database Systems for Advanced Applications), pp.247-254, 1996.
2. T.Matsui and S.Furui: " Comparison of text independent speaker recognition methods using VQ distortion and discrete/continuous HMMs", Proc.ICASSP, Vol.II, pp157-160, 1992.
3. Y.Ariki and K.Doi, " Speaker Recognition based on Subspace Method", ICSLP'94, pp.1859-1862, 1994.
4. E.Oja,"Subspace Methods of Pattern Recognition, Research Studies Press, England, 1983.
5. Cambridge University Engineering Department Speech Group and Entropic Research Laboratory Inc.:"HTK Hidden Markov Model Toolkit V2.0"
6. Y.Ariki and Y.Saito : " Extraction of TV News Articles based on Scene Cut Detection", ICIP96, pp.III847-III850, 1996.