AN ENQUIRING SYSTEM OF UNKNOWN WORDS IN TV NEWS BY SPONTANEOUS REPETITION

(Application of Speaker Normalization by Speaker Subspace Projection)

Y.Ariki and S.Tagashira

Department of Electronics and Informatics
Ryukoku University, Seta, Otsu-shi, 520-21, Japan
ariki@rins.ryukoku.ac.jp

ABSTRACT

We tried to construct a system in which we can enquire unknown words, included in TV news speech by repeating them spontaneously. For example, we hear "Japan would join PKO." from TV news and if "PKO" is an unknown word, then we can enquire it by saying "What's the PKO?" The system recognizes the word "PKO" and explains its meaning. In this system, it estimates a common section between announcer's speech and user speech, and recognizes the word corresponding to the common section. We solved a problem of speaker difference in extracting common sections by speaker subspace projection.

1. INTRODUCTION

TV news program broadcasts Multi-media information composed of characters, images and speech. When we meet with interesting but unknown words in the program, we would like to know them by consulting with a dictionary or asking a person. But it is troublesome to consult with a dictionary, even if it is electronized and we could type a keyboard. What We need is a voice activated enquiring system instead of keyboard, because it is natural to enquire unknown words by voice when we are getting TV news information via voice media. From this point, we tried to construct this kind of enquiring system by repeating the unknown words spontaneously just after we hear it from announcer's speech.

Fig.1 shows a conceptual image of the system. We can enquire unknown words such as "stagflation" in a spontaneous manner like "Oh, what is the stagflation, please?" , "Stagflation what?" or "Tell me the stagflation" , when we hear ".... economical growing under stagflation is ....." from TV news. The system recognizes the word "stagflation" , because it is commonly included in the user speech and announcer's speech. Then the information of the word "stagflation" is retrieved from a CD-ROM of an electronized economic dictionary and presented to the user.

Spontaneous speaking style is inevitable because the user has no time to speak long rigid sentences just after hearing them. In most cases, we speak one word or an abbreviated short sentence such as "stagflation?" , "Oh, what's the stagflation?" , "Stagflation what?" or "Tell me the stagflation" . In order to recognize enquired words, keyword spotting seems
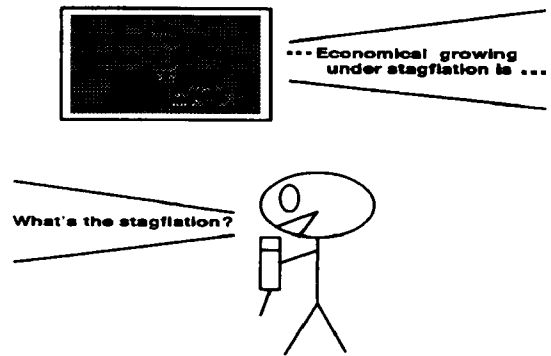

Figure 1: Conceptual image of a system

to be powerful[1]. However it is difficult to spot the word independently both on the user speech and announcer's speech, when a dictionary size becomes enormous.

To solve this problem, we employ a method to extract common section between the user speech and announcer's speech, and then to recognize the word in the common section. As an extraction method of the common section, we employ Reference Interval-free Continuous DP (RIFCDP) proposed by Oka[2]. This method is the extension of continuous DP with end-ponts free on both two utterances. After applying this method, the common section is extracted as the partial section with the highest score between them. However this extraction method has a problem to be sensitive to speaker difference as is the DP matching.

To solve this problem, we employed a speaker normalization method which projects speech data to individual speaker subspace presenting speaker characteristics[3]. Announcer's subspace is independently produced at first, and then the user subspace is produced so as to maximize the correlation of the axes between the subspaces of the announcer and user.

2. SYSTEM ORGANIZATION

The system consists of following 5 blocks as shown in Fig.2.

(a) SPEECH INPUT

Announcer's speech is digitized into a circular buffer to keep the latest utterance. User speech is always watched and if an user speaks, the system begins to digitize his voice.

(b) SPEAKER NORMALIZATION

User voice is normalized (adapted) to announcer's voice. Announcer's subspace is constructed at first and then user subspace is constructed so as to maximize the axes correlation between the user's subspace and an announcer's subspace. The normalized speech data are presented in the local coordinates of the respective subspace.

(c) WORD EXTRACTION

Dynamic Programming is performed on the normalized speech data between the announcer and the user in order to extract the time sections commonly included in both speech. For this extraction, Reference Interval-Free Continuous DP (RIFCDP) proposed by Oka is used.

(d) WORD RECOGNITION

Word recognition is applied using HMMs to all the sections extracted in (c). The section with the highest probability is confirmed as the word.

(e) INFORMATION RETRIEVAL

The recognized word is sent to information database with CD-ROM and the associated text is retrieved and synthesized by speech.
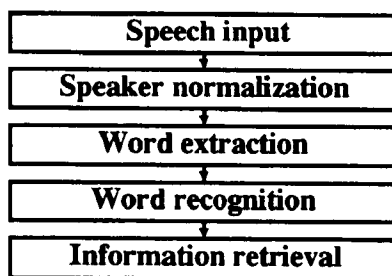
```
┌─────────────────────────────┐
│      Speech input           │
└─────────────────────────────┘
             ↓
┌─────────────────────────────┐
│   Speaker normalization     │
└─────────────────────────────┘
             ↓
┌─────────────────────────────┐
│     Word extraction         │
└─────────────────────────────┘
             ↓
┌─────────────────────────────┐
│     Word recognition        │
└─────────────────────────────┘
             ↓
┌─────────────────────────────┐
│   Information retrieval      │
└─────────────────────────────┘
```

Figure 2: System organization

3. SPEECH INPUT

In the system, TV news are already digitized as digital videos and we can playback them on a computer. User speech is always watched and if an user speaks, the system begins to digitize his speech. Announcer's digital speech is taken out for 5 seconds just before user speaks under the assumption that the user begins to enquire within 5 seconds after he hears an unknown word. Table1 shows the condition of speech analysis used in the system.

Table 1: Condition of speech analysis

Sampling frequency	12kHz
High-pass filter	$1 - 0.97z^{-1}$
Feature parameters	LPC cepstrum(16th)
Frame length	20ms
Frame shift	5ms
Window type	Hamming window

4. SPEAKER NORMALIZATION

4.1. Decomposition of Speech Data

As shown in Fig.3, we observe speech data X_A of speaker A and speech data X_B of speaker B in an observation space. The speech data are a sequence of spectral feature vectors x_{At} and x_{Bt} obtained at time t by short time spectral analysis. We denote the speech data X_A as a matrix whose row is a spectral feature vector x_{At}^T, $(1 \leq t \leq M)$. The column of the matrix corresponds to frequency i, $(1 \leq i \leq N)$.

By singular value decomposition, the speech data matrix X_A is decomposed as

$$X_A = U_A \Sigma_A V_A^T \tag{1}$$

Here U_A and V_A are the matrices whose columns are eigenvectors of $X_A X_A^T$ and $X_A^T X_A$ respectively, and Σ_A is the singular value matrix of X_A.
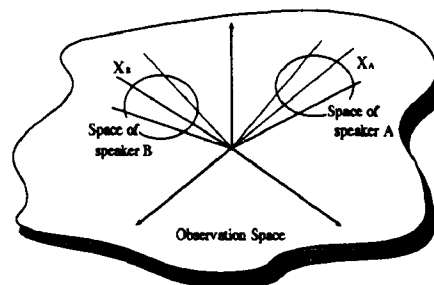

Figure 3: Observation space and speaker space

4.2. Speaker Subspace and Normalized Data

The eigenvectors of the correlation matrix $X_A^T X_A$ are the orthonormal bases of the speech data X_A, computed based on a criterion that the total distance is minimized between observed speech vectors x_{At} and the orthonormal bases.[3] Then V_A is considered as orthonormal bases of the speaker space.

If the large r singular values are selected from the matrix Σ_A, the matrix U_A becomes M x r dimension and the row still corresponds to time. The matrix V_A^T becomes r x N dimension and is considered as speaker subspace.

Since the speech data matrix $U_A\Sigma_A$ is presented in own speaker subspace, we can say that speaker characteristics is less included in $U_A\Sigma_A$ than the speech data matrix X_A presented in the observation space. This interpretation indicates that $U_A\Sigma_A$ is the speaker normalized data and has mainly phonetic information. On the other hand, V_A is the speaker subspace and has mainly speaker characteristics. Speaker normalization is realized using the speaker normalized speech data $U_A\Sigma_A$.

4.3. CLAFIC Canonical Correlation

The simplest method of speaker normalization is CLAFIC method[4] which computes announcer subspace (orthonormal bases) V_A by the singular value decomposition at first using speech data X_A spoken by the announcer. The speech data X_A are projected to the announcer subspace from the observation space, and the speaker normalized data $U_A\Sigma_A$ are obtained. In the same way, if speech data X_B of the user are given, the user subspace V_B is independently computed and the speaker normalized data $U_B\Sigma_B$ are obtained. The speaker normalized data $U_A\Sigma_A$ and $U_B\Sigma_B$ are used for extracting common section.

This CLAFIC method has a problem that correlation of the subspace between the announcer and the user is not considered. The canonical correlation analysis [5] is well known to solve this problem. However, the canonical correlation method has another problem that the subspace produced by the canonical correlation analysis does not present the speech data discriminatively. It also causes the problem that the $U_A\Sigma_A$ changes when another user C is normalized, because the subspaces of the announcer A and the user B are simultaneously produced by the canonical correlation analysis. To solve these two problems, we propose a CLAFIC canonical correlation analysis in which the subspace of the announcer A is independently produced by the CLAFIC method at first, and then the subspace of the user B is produced as to maximize the correlation of the axes between the subspaces of the speaker A and B. The step is summarized as follows;

STEP(1) Feature vectors in spoken sentences are matched by DP between the announcer A and the user B, and the matched speech data X_A and X_B are obtained.

STEP(2) Orthonormal bases V_A of the announcer A are computed using the speech data X_A in the same way as the CLAFIC method.

STEP(3) The axis v_B of the user B is computed as follows in the way of maximizing the correlation between the axes v_A and v_B using the speech data X_B.

$$v_B = \frac{\sqrt{C}\Sigma_{22}^{-1}\Sigma_{21}v_A}{\sqrt{v_A^T\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}v_A}} \qquad (2)$$

where C is the variance on the axis v_A. Σ_{22}, Σ_{21} and Σ_{12} are auto-correlation and cross-correlation matrices respectively.

4.4. Speaker normalization result

As a preliminary experiment, we carried out 50 words recognition between two speakers after speaker normalization using 162 words. The speech database used is spoken word database produced by Tohoku University and Matsushita cooperation. We selected a male speaker sp210 as a standard speaker to prepare 50 word templates. Two other speakers, sp301 (male) and sp606 (female) were selected as input speakers for 50 word recognition. Table2 shows the recognition result. The recognition rate was improved from 84.0% to 94.0% for male speaker and from 62.0% to 92.0% for female speaker. The subspace dimension was set to 16 in this experiment. From the table, it can be said that the speaker normalization is effective because 30% improvement was obtained between male and female speakers.

Table 2: Word recognition result by DP(%)

	Male(sp301)	Female(sp606)
Without normalization	84.0	62.0
With normalization	94.0	92.0

5. EXTRACTION OF COMMON SECTION

Common sections are extracted between announcer's speech and user speech after speaker normalization. As an extraction method, we employed Reference Interval-free Continuous DP (RIFCDP) proposed by Oka[2]. This method is the extension of continuous DP with end-ponts free on both two utterances. After applying this method, the common sections are extracted as the partial section with higher score between them.

Fig.4 shows a concept of RIFCDP proposed by Oka[2]. The τ denotes the τ-th frame in the announcer's speech. Matching score $(M_{\tau_3}^{\tau_1})$ indicates the best DP matching score between the section in the announcer's speech from frame τ_1 to frame τ_3 and the section in the user speech ending at frame t. In computing the matching score $(M_{\tau_3}^{\tau_1})$, the accumulated distance $(D_{\tau_3}^{\tau_3})$ to frame τ_3 at frame t and the best path should be computed at first and then the accumulated distance $(D_{\tau_3}^{\tau_1})$ to frame τ_1 at frame t_3 on the best path is subtracted. In order to make it easier, a series of accumulated distance on the best path is stored at the arbitrary frame on the announcer's speech. For example, a series of accumulated distance $D_{\tau_3}^{\tau}$ $(1 \leq \tau \leq \tau_3)$ on the best path to frame τ_3 at frame t is stored on the frame τ_3. The matching score $(M_{\tau_3}^{\tau_1})$ at frame t can be computed by subtracting the stored accumulated distance. The matching score is computed for all input frames t and for all frames on announcer's speech and then the common section is extracted by local peak picking or thresholding of the matching score.

6. WORD RECOGNITION AND INFORMATION RETRIEVAL

Word recognition is carried out on all the extracted common sections using speaker independent HMMs. Phoneme
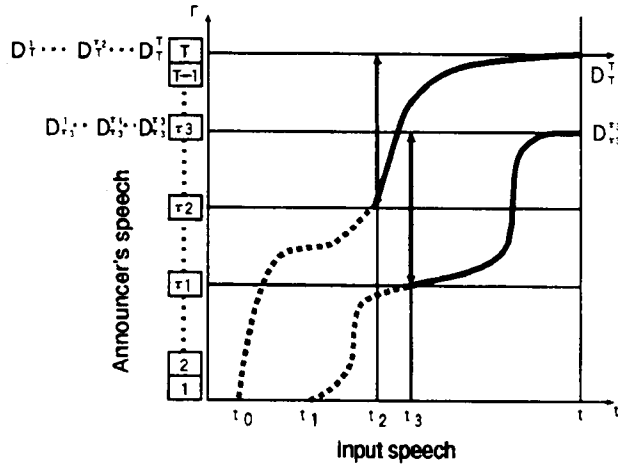
Figure 4: Concept of RIFCDP

HMMs are concatenated according to a dictionary and word HMMs are constructed. Table3 shows the recognition condition. The number of speakers are 50 and 10 spoken sentences are used from each speaker for HMM training.

Since several common sections are extracted by RIFCDP between two utterances, the common section with the highest probability by HMM word recognition is confirmed as the target word. The recognized word is sent to information database with CD-ROM and the associated text is retrieved and synthesized by speech.

Table 3: Recognition condition

HMM type	Mixture Gaussian HMM
Number of states	5 states 3 loops
Number of mixtures	4
Covariance matrix	Diagonal matrix
Speaker	Independent

7. SYSTEM EVALUATION

7.1. Experimental Condition

We evaluated the total system performance through speech input, speaker normalization, word extraction, word recognition and information retrieval. The announcer's speech was collected from 30 days NHK TV news each of which was broadcasted for 5 minutes. In speaker normalization, 20 sentences were selected from the announcer and the user, and used for CLAFIC canonical correlation analysis. The condition of speech analysis was same as shown in Table1. At present, 25 words are prepared as a small dictionary for word recognition. According to these 25 words, 25 sentences were selected from announcer's speech. User speech of 25 short sentences spoken by one male were also collected. The total system performance was evaluated as a ratio of correctly retrieved number of words to 25 trials.

7.2. Experimental Result

Table4 shows the result of information retrieval experiment carried out as the total system evaluation. The retrieval rate is improved from 72.0% to 80.0% by normalization. The errors are classified into following 3 types.

(1) Extraction of common section is wrong.

(2) Extraction is correct but Recognition is wrong.

(3) Extraction is correct but rejected because other common section has higher recognition probability.

(1) indicates there still exist extraction error of common sections after speaker normalization. The threshold of RIFCDP should be more investigated. (2) occurred between similar words so that this kind of error will be reduced as well as (3) by improving the HMM accuracy.

Table 4: Result of system evaluation (%)

	Retrieval rate
Without normalization	72.0
With normalization	80.0

8. CONCLUSION

We have tried to construct the system which can retrieve unknown words enquired by a user in speech just after he hears it in the TV news program. This system can accept spontaneous speech which includes unknown word to him. To extract the common section between announcer's speech and user speech, we employed RIFCDP. We also employed CLAFIC canonical correlation analysis for speaker normalization to solve the problem that RIFCDP is sensitive to speaker difference.

9. REFERENCES

1. P.Jeanrenaud,K.Ng,M.Siu,J.R.Rohlicek,H.Gish: "Phonetic-Based Word Spotter:Various Configurations and Application to Event Spotting" , Proc. ESCA EuroSpeech93, pp.1057-1060, 1993.

2. Y.Itoh,J.Kiyama,H.Kojima,S.Seki,R.Oka: "Reference Interval-free Continuous Dynamic Programming (RIFCDP) for spotting speech waves by arbitrary parts of a reference pattern" , Technical report of IEICE, SP95-34, 1995.

3. Y.Ariki and K.Doi, : "Speaker Recognition based on Subspace Methods" , ICSLP94, pp.1859-pp.1862, 1994.

4. E.Oja, : "Subspace Methods of Pattern Recognition" , Research Studies Press, England, 1983.

5. K.Choukri G.Chollet Y.Grenier : "Spectral transformations through Canonical Correlation Analysis for speaker adaptation in ASR" , ICASSP86, pp.2659-2662, 1986.