Research on the Optimal Strategies of Chinese News Program Catalog based on ASR

Lijin Long[1, 2], Yuyou Cheng[2]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, 310027, China

[2]Department of Electronic and Information Engineering, Zhejiang University of Media and Communications, Hangzhou, 310018, China

email: zj_education_ljl@163.com

Keywords: News Program Catalog; ASR; Optimal Strategy; meta-data; STR

Abstract. In the process of generation of meta-data, speech in news programs can be translate directly into text information based on ASR technology so as to improve cataloger's efficiency and the quality of meta-data before being put in media asset management database. For the purpose of this application, four optimal strategies has been put forward to improve the accuracy of text information extraction from news programs. The experimental results indicate that this four optimal strategies are efficient in extracting text information and the developed software based on the four optimal strategies has been put into catalog for news programs.

Introduction

In recent years, with the development of the hardware and software technology, automatic speech recognition (hereafter referred to as ASR) is becoming more and more widely applied in various of fields, which include real time voice input, voice command control, multimedia audio information retrieval, information monitoring, Video content behavior recognition, mutual translation between speech and text, interactive voice response (IVR), and so forth. Due to speech in news programs has many advantages from speech to text based on ASR, such as newscaster has good mandarin level, news program is easily to be segmented into one or more sections, speech recognition training is easier for newscasters, and so on. The research and application of speech recognition is gaining more and more attention in news programs[1].

In media asset management system, or simply MAMS for short, the material of video and audio need to be cataloged into meta-database by catalogers before being put into meta-database so that massive amounts of programs in MAMS can be convenient to manage by meta-data and to improve the utilization rate of program resources in meta-database. In the process of generating meta-data cataloged from programs, especially for news programs, a large number of speech information can be directly translated into entries of audio material meta-data, which can be used for reference to cataloging related meta-data into meta-database. Therefore, the application of automatic speech recognition technology in news program catalog can improve not only the catalog efficiency but also the quality of news program meta-data.

The main challenges to automatically extract text information from speech of news programs by software are preprocessing of hybrid speech signal, the selection of processing length for speech signal, the personalized voice features of news casters, such as intonation, speed, rhythm, accent, pronunciation habits, etc.

In this paper four optimal strategies are put forward to extract text information from news programs based on ASR. Its contents and structures are as follows. First, this paper introduces the background and significance of applying ASR into news program catalog. Second, the characteristics of news programs are analyzed and the procedure how to apply ASR into the process of news program catalog is introduced. Third, it puts forward four optimal strategies improve the accuracy of text information extraction. At last, it discusses the prospects about the application of ASR and concludes the advantages of the proposed approach for text information extraction in news programs.

Application of ASR in News Program Catalog

Before ASR is applied into text information extraction for news programs, in view of the internal working principle of speech recognition engine[2,3,4], news programs can be preprocessed first based on their features, such as newscasters' mandarin level, news programs can be segmented into one or more sections, newscasters' speech can be trained, etc. After that, based on speech recognition engine, the performance of text information extraction will be improved by the developed software for news programs so that the efficiency that catalogers apply these text information into generating meta-data is increased and the quality of meta-data is also improved.

Analysis of Text Information Extraction for News Programs.

In ASR system, speech recognition engine has a great influence on the accuracy of text information extraction from speech of news programs. The key step of speech recognition engine is pattern matching. Currently, the main algorithms applied into speech recognition engine include DTW, HMM, VQ, ANN, etc. Among these algorithms, HMM and ANN are the two most widely used. In consideration of the fact, newscasters' mandarin level and accent are better than other ordinary people, in the process of pattern matching, the recognition results have less error prone, which is one of the advantages to the goal of this paper.

The speech recognition engine mainly deals with two kinds of data[2,3]. The one is model data adopted by the speech recognition algorithm, such as acoustic layer, language layer, spelling grammar, etc. The other is user data generated by feature database, such as user profiles, phonetic characteristics, corresponding relationship of recognition model, etc. In TV stations and radio stations, the newscasters are generally stable, therefore, before applying ASR into extracting text information from news programs, some speech training can be made according to newscasters' characteristics for specific selected news programs to generate user profiles, which can improve the recognition accuracy in text information extraction.

In the process of extracting text information from news programs by speech recognition engine, the consecutive length of voice have a great influence on recognition accuracy. On the one hand, news programs consists of one or more news sections, speech of news sections announced by the same newscaster generally. On the other hand, the personnel speed of newscasters’ broadcast is relatively quickly. These two features make continuous speech recognition module in speech recognition engine to misinterpret several statements for a whole statement and recognize them by associative memory processing, which increase the recognition difficulty and lead up to less recognition accuracy. By means of dynamic segmentation and reduction of speech speed in news programs can ameliorate the above-mentioned problem.

Procedure of Text Information Extraction.

At present, the practical application systems about speech recognition software mainly include the IBM Via Voice, Inter-phonic series, the voice component internally installed in Microsoft Windows XP, Windows 7, or Windows 8, etc., which can also develop the voice recognition software by their provided SDK interface. These speech recognition engines have progressed gradually and steadily. In addition, there are some applications developed specifically for some areas of expertise such as robot control, search engine based on the content, direct conversion from audio to text(Wave to Text - Voice Recognition), etc. For news programs, there are two kinds of methods to reach the goal of extracting text information from speech in news programs. The one is to apply the developed software provided by some major corporations, which engage in speech recognition research, into conversion from voice to text directly. The other is to develop for special application by SDK packages based on some preprocessing. According to experiments, the paper concludes to directly apply the software into extracting text information from news programs mentioned above without preprocessing has lower accuracy than it does, which results from the internal speech recognition engine of software have no good optimization for Chinese and the main task of internal speech recognition engine is not just to extract but to recognize. Based on the above analysis, the paper will follow by the steps how to apply the speech recognition engine into developing software for automatic extracting text information from news programs based on Microsoft Speech SDK.

1. Do some preprocessing for news programs (see part 3 in this article).

2. Set up the record of PC sound card with mixing style.

For news programs, there exists some problems caused by electromagnetic noise (also called electromagnetic frequency EMF) and background noise, which have brought a great obstruction to the accuracy of news speech recognition. Therefore, the paper suggests that the speech input adopts to internal record way for news programs.

3. Open the software developed based on Microsoft Speech SDK.

The paper use Microsoft Speech SDK to generate the corresponding text information automatically. The built-in voice recognition software in Microsoft's operating systems can be also used to extract text information from news speech manually. Generally, the more advanced the OS is, the better the accuracy is.

4. Generate corresponding text information automatically from news programs.

The developed software, which based on speech SDK packages or internal components provided by OS, can generate corresponding text information automatically from news programs as a reference for catalogers to put meta-data into meta-database after some preprocessing steps have been done. In view of text information extracted based on news speech, if the news program consists of two parts, the one is video and the other is audio, there is no need to open the video because audio can reach the goal only. For some news programs, if media player applied to extract text information does not support speech code, the news programs need to be transcoded before processing.

As shown in Fig.1, the result of direct extraction of text information from CCTV news section has lower accuracy and has more difficult to be used by catalogers. In part “Test Analysis and Conclusion”, the paper will show the better result after some preprocessing steps have been done before extracting text information from news programs.

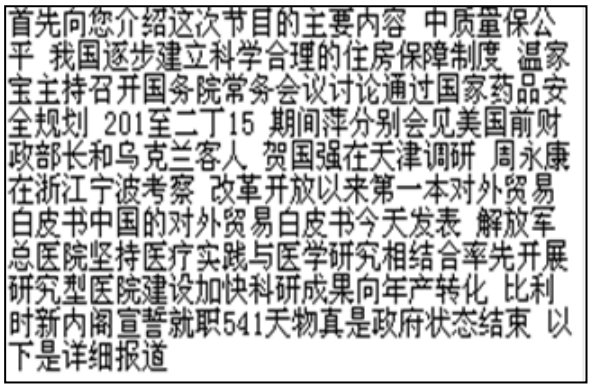

Fig.1.The experimental result extracted from 2011.12.07 CCTV news

Improvement of Text Information Extraction for News Programs

The challenges to extract text information from news programs mainly include mixed voice signal, the length of speech signal (especially for that amplitude is higher than threshold value of continuous time), personalized voice features of newscasts, etc[5,6,7]. For news programs, the paper try to overcome these challenges so as to improve recognition performance by preprocessing, which include speech purified, cutting news programs into sections, reducing of speech speed, voice training, etc.

News Speech Purification.

The process of news speech purification is close related to the process of blind signal separation(BSS)[8]. Given a source signal, voice signal of newscaster can be estimated by solving the separated matrix W(see Eq.3). The main challenges existed in BSS come from the limitation of recovery criteria as well as the lack of prior knowledge. The method BSS can only get waveform of source signal and it can't determine the amplitude value and order of signal. In the process of BSS, there need to assume that statistical independence, linear mixture, no more than one Gaussian signal

in source signals because they can't be separated when there has superposed Gaussian signal in source signals. Formal description of BSS as follows. Suppose source signal vector is:

$$S(t) = [s_1(t), s_2(t)...s_n(t)]^T \quad (1)$$

In Eq.1, there has the statistical independence among the components. The source signal vector S is multiplied by a mixing matrix A(m*n) for observation vector X. The predicted value vector Y can get by separated matrix W multiplied by source signal vector X.

$$X(t) = AS(t) = [x_1(t), x_2(t)...x_m(t)]^T \quad (2)$$

$$Y(t) = WX(t) \quad (3)$$

In order to measure the performance indicator of W, crosstalk error(ETC) is defined as follows:

$$E = \sum_{i=1}^{n}(\sum_{j=1}^{n}\frac{|c_{ij}|}{\max_k|c_{ik}|}-1) + \sum_{j=1}^{n}(\sum_{i=1}^{n}\frac{|c_{ij}|}{\max_k|c_{kj}|}-1) \quad (4)$$

In Eq.4, variable C_{ij} represents the element of row i and column j in matrix $C = WA$. Generally, the smaller the variable E is,the better the result is. The lower limit of variable E is zero.

In order to solve matrix W, there have several methods such as ICA, EASI(LMS), natural gradient RLS, natural gradient RLS, and so on[9,10], which can be classified into two categories, they are the method based on statistics and the method based on forecast. By applying these algorithms into news programs for purifying news speech, we found EASI achieves best result due to the reason that news program belongs to non-balance signals although EASI has lower speed of convergence than the others. Table 1 shows the ETC weighted average values result from applying different algorithms into purifying news speech for about 50 news videos with background noise in experiments. The weights of news programs are decided by their time length.

Table 1 Different ETC values come from different algorithms

Algorithm	ICA	EASI	RLS	Natural Gradient RLS
ETC	3.79	0.21	0.63	2.88

These nearly 50 news videos selected from news program library are mainly divided into four categories: 1.The videos with mixed voices based on the fact voice of newscast and news content is playing in order. 2.The videos with multiple voices based on the fact newscast's voice is higher than the background chaotic noise[11]. 3.The videos with some sections based on the fact the news content is switching to another news section. 4.The videos with sudden undesired sounds such as the sound of strong winds, the sound of rain, the sound of sea, etc. Based on experiments on news programs with different algorithms, the paper concludes the main reason that EASI shows better effect for news programs is that the other algorithms such as ICA,RLS, etc. depend on the statistics of signals more than EASI does.

Implementation of Auto-segment for News Programs.

From the observation on the signal of speech, news programs are continuous signals in time domain. As shown in Fig.2, the high oscillation frequency existed in speech signals represent the news sections of announcing and the flat part represents rest, switch or faint noise. When a piece of news has finished or in the process of switching to another news section, there exists a period of flat generally and oscillation comes again after that. According to the above analysis, one news program can be divided into one or more than one segments according to the duration of the flat amplitude based on the analyzing on time domain. In order to achieve better result, there need to solve two key problems: setting of threshold value and length of consecutive idle period. The purpose of setting of threshold value is to normalize those small sampling points to zero so as to reduce the interference of those points to segment and the setting of consecutive idle period length is the basis of segmentation so that the length of short idle period should be ignored in case of over-segmentation.

In our experiments, we made some statistics based on nearly 50 news programs. The result shows the threshold value is set to 0.01~0.02 and the length of consecutive idle period is set to about 2000 sampling points can achieve comparatively better segmentation accuracy for most of

news programs selected from news program library.

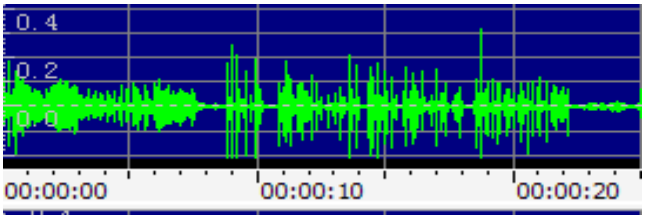

Fig.2. The phonetic signal of news program

Reduction of Speech Speed.

With the purpose of achieve good result for extracting text information from news programs, one of strategies, the paper put forward, is to slow down speech of news programs appropriately. Inevitably, the slower the voice is, the greater distortion the voice is. Therefore, we need select an appropriate algorithm for this kind of special task. By experiments, a single extension on time domain of speech will lead to deformation of voice and some interpolation algorithms appear to unstable performances on recognition accuracy, such as linear interpolation, cubic spline interpolation, two dimensional cubic convolution interpolation, and so on. In contrast, after the procession of changing frequency and interpolating with FFT to slow down the speech of news programs can improve recognition accuracy to a certain extent.

As shown in Fig.3, the signal located in the top half is source signal of news program and the signal located int the bottom half is processed signal by FFT with extension of 1/4 in time domain. From the effect of audition, the processed signal has small distortion and the speed of voice has slowed down to be appropriate for improving the accuracy of text information extraction on news programs.

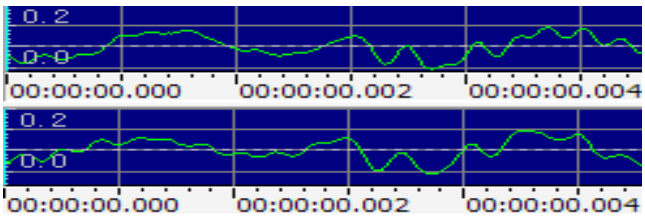

Fig.3. The comparison between original speech signal and preprocessed speech signal with time domain extension

Speech Recognition Training.

Generally, ASR system provides user interface to improve performances of speech recognition engine by training so that speech recognition engine can adapt to different application environments. It's difficult for ASR to recognize isolated words, unfamiliar words, semantic incoherent words, such as Chinese names, place names, country names, rarely used words, loan words, network hot words, etc. For news programs, by speech recognition training, on one hand, the speech recognition engine can recognize the special vocabularies of the news, for example common used Chinese names(such as national leaders, relative to a particular event, famous people, etc.), national names(diplomatic meetings, economic and military development, etc.), place names(such as capital, provincial capital, state name,etc.). On the other hand, the speech recognition engine can improve adaptability for phonetic characteristics of newscaster. The specialized vocabularies of news programs can be summarized as several of lists for different kinds of news programs according to commonly used words. The phonetic characteristics and personalized training can also be corrected by the recognition effect of news real-time content, which results in "user profile" to improve recognition accuracy. The specific procedures, how to train speech recognition engine, are as follows based on speech recognition engine in Microsoft operating system.

The ASR built-in Microsoft operating system can use training card to correct ASR errors happened during training. It also provides user interface to expand terminology vocabularies so as to improve recognition performance for special words, consecutive semantic phrases and sentences.

As to the correspondence between standard speech and phonetic features, the ASR built-in Microsoft OS doesn't provide the interface to change training samples so there are two methods to

improve the speech recognition engine by training. The one is to backup user profile by built-in tool for the speech recognition engine improved by training. In practical application, after the same sound sample hardware environment is built, the backed up user profile can be imported in ASR. The other is to correct those error recognition content based on specific training by catalogers. For the latter, after the speech recognition engine improves recognition rate for training voice content, although different sound devices in PC have different sample features, it improved recognition accuracy more or less in experiments. In contrast, the former method is better.

Test Analysis and Conclusion.

To test the method proposed in this paper, based on Microsoft Speech SDK, we developed a application integrated with above -mention four optimal strategies for nearly 50 news programs. Due to the difference among news programs such as mandarin level, speech speed, speech environment, etc., automatic text information extraction from news programs vacillated in a range of 30%-85% in statistics before applying above-mention four optimal strategies into preprocessing. According to analyze on different news programs, we applied one or more than one strategy into these experimental samples, the whole recognition accuracy improve a lot in recognition accuracy to reach above 85% for most news programs in experiments. We did some special training for some news programs to train purposeful training of unusual words, conjunctions, the consecutive announce habit of newscasters, the automatic transition from voice to text can improve recognition accuracy to over 98%, which can not only bring convenience to catalogers and provide a lot important information for catalogers to make meta-data relative to news programs better used for management in practical application.

As shown in Table 2, based on different news categories, we classified about 50 news programs(split into 252 news sections) into 5 classes and compared the recognition rate of text information extraction between original videos and processed videos with above-mention four optimal categories. The results show the proposed approach to extract text information from news programs in the paper has a great improvement in contrast to original news programs. One of experimental results of text information extraction directly from processed news sections selected from 252 news section library is shown in Fig.4.

Table 2 The experimental results for news programs

News Program Type	News Sections	Accuracy (%)	Accuracy after preprocessing(%)
Briefing News	47	40-85	95-98
Live Report	23	30-75	85-95
Interview	57	40-80	85-98
Commentary	29	35-80	85-95
Others	96	30-85	85-98
Summary	252	30-85	85-98

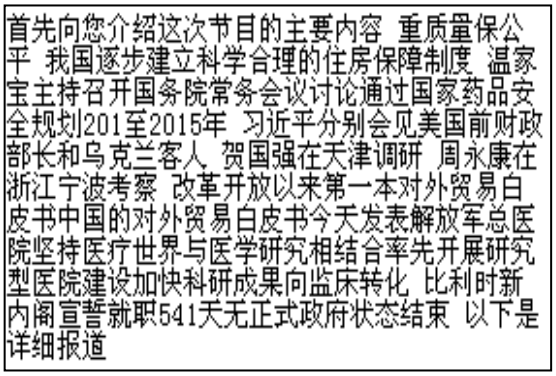

Fig.4. Text information extraction form CCTV news

Conclusion

In recent years, the application of automatic speech recognition technology is becoming increasingly wider and deeper, such as tourism, banking, social services, and other fields [12,13], which also promote the chip industry of ASR, for example, Pod Zinger search engine developed by Every Zing can translate directly voice into searchable text. According to experimental results, its recognition rate can reach up to 80% or more in pure voice environment(single voice signal). Moreover, after applying deep neural network technology into speech recognition engine, the recognition rate of speech recognition engine developed by Microsoft will raise from 50% to 80% or more.

For the purpose application, text information extraction from news speech based on ASR, this paper puts forward four optimal strategies to improve recognition rate for news programs. The developed software based on the proposed approach and designed for news catalog has been applied to improve the work efficiency of catalog as well as promote the quality of meta-data before being put in meta-database.

References

[1] ChaoKyang C U I. The present situation of speech recognition and the application in the field of radio and television. Computer Engineering and Applications, 2007, 43(23).

[2] Lee A, Kawahara T. Recent development of open-source speech recognition engine julius[C]//Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference. Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, 2009: 131-137.

[3] LIU Y, LIN Y, CHEN Z. Text Information Extraction Based on Hidden Markov Model [J]. Acta Simulata Systematica Sinica, 2004, 3: 038.

[4] Kumar K, Liu C, Gong Y. Normalization of ASR Confidence Classifier Scores via Confidence Mapping[C]//Fifteenth Annual Conference of the International Speech Communication Association. 2014.

[5] Huang Y M, Liu C J, Shadiev R, et al. Investigating an application of speech-to-text recognition: a study on visual attention and learning behavior[J]. Journal of Computer Assisted Learning, 2015.

[6] Do C T, Lamel L, Gauvain J L. Speech-to-Text Development for Slovak, a Low-Resourced Language[C]//Spoken Language Technologies for Under-Resourced Languages. 2014.

[7] Vasilescu I, Vieru B, Lamel L. Exploring pronunciation variants for Romanian speech-to-text transcription[C] //Spoken Language Technologies for Under-Resourced Languages. 2014.

[8] Smith D, Lukasiak J, Burnett I. Blind speech separation using a joint model of speech production[J]. Signal Processing Letters, IEEE, 2005, 12(11): 784-787.

[9] Lee J H, Jung H Y, Lee T W, et al. Speech feature extraction using independent component analysis[C]//Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on. IEEE, 2000, 3: 1631-1634.

[10] Tokuda K, Kobayashi T, Imai S. Speech parameter generation from HMM using dynamic features[C]//Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on. IEEE, 1995, 1: 660-663.

[11] Schmid D, Thüne P, Kolossa D, et al. Dereverberation preprocessing and training data adjustments for robust speech recognition in reverberant environments[C] //Speech Communication; 10. ITG Symposium; Proceedings of. VDE, 2012: 1-4.

[12] Gemignani G, Bastianelli E, Nardi D. Teaching robots parametrized executable plans through spoken interaction[C]. AAMAS, 2015.

[13] Jones G J F. Speech search: techniques and tools for spoken content retrieval[C]//Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014: 1287-1287.