# Key Information Extraction from Broadcast in Traffic Domain

J.B.Gu, Y.F.Xie, Y.Y.Wang, X.Xie
Beijing Institute of Technology
Beijing, China

***Abstract* — With the development of Artificial Intelligence, Automatic Speech Recognition has become an important field of cognitive computing in terms of intelligent vehicle. This paper introduces the application of Speech Recognition and Natural Language Processing on intelligent vehicle that extracts key information from traffic broadcast. This system consists of three modules, speech detection module, keywords output module, and structured information extraction module. The speech detection module uses Non-negative matrix factorization to make speech enhancement, which separates the speech from the background music. Using Mel-frequency Cepstral Coefficients (MFCCs) extracted from the speech as features, it trains the model using Hidden Markov Model (HMM). Language model like n-gram improves the performance of the recognizer. In structured information extraction module, Conditional random field (CRF) is used to make partial parsing. Then it fills the slots, and outputs structured traffic condition information. We give the experiment results of every module.**

***Keywords-automatic speech recognition; artificial intelligence; natural language processing; HMM; CRF***

## I. INTRODUCTION

Unmanned Ground Vehicles (UGV) has become a hot spot in both academy and industry. It uses camera, radar and other sensors to sense the environment information that guides it to make path decisions. To gather more information, Speech Recognition are necessary to extract traffic condition information from traffic broadcast, which aids vehicle to make path decision in a long distance. The system should detect the broadcast, convert it to text, and extract information that can be used by modules in the back end. So the whole system contains following modules, shown in Figure 1.

(1) Speech detection module.

The system records the broadcast speech, then split the audio signal by End point Detection. It pre-process the signal to filter the noise.

(2) Keywords output module.

This module extracts MFCC features and trains HMM models using corpora. To increase the accuracy, n-gram model together with the dictionary are built.

(3) Structured information extraction module.

To the keywords sequence, it makes partial parsing. Then it fill the slots. At last, it outputs structured information.

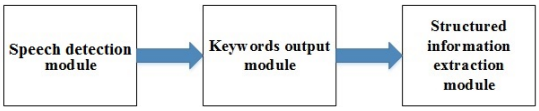


FIGURE I. THE WHOLE SYSTEM FRAMEWORK

## II. CORPORA

Corpora are important in training recognizer. To build the system, audio corpora are collected from Beijing Traffic Broadcast since June 2013 to December 2013. All the audio files are up to 36 hours. People are recruited to make annotations. Their task is to convert the speech to text and to make some annotations manually, including making sentence boundary annotations, slips of tongue annotations, repetition annotations and breathe annotations.

These corpora contain not only traffic condition information, but also speakers' talk show, advertisement description, and much other useless information, which is language level noise. Also, it contains strong background music, which is acoustics level noise. It contains about 20000 sentences, 7932 of which are effective sentences in traffic domain. What should be focus on is the traffic condition information. It should be output in a structured form.

Besides, written language corpora in traffic domain are crawled from Microblog and Beijing Traffic Information website. They aims to train language model.

## III. SPEECH RECOGNITION SYSTEM

The speech recognition system is shown in Figure 2. The recognizer consists of acoustic model, language model and recognition network.

### *A. Pre-Processing*

The original speech signal should be processed by a pre-processor to filter noise. As much noise exists in our corpora, we use speech enhancement approach based on NMF to denoise the signal. That is to separate speech from music. The algorithm can be seen in [1].

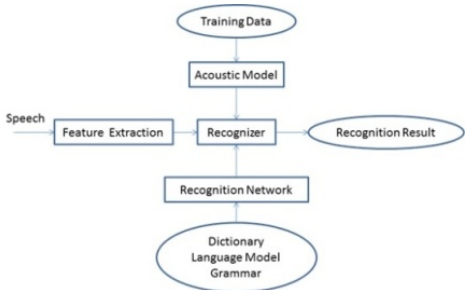


FIGURE II. SYSTEM FRAMEWORK OF SPEECH RECOGNIZER

### B. *Features Extraction*

To train recognition models, acoustic features are extracted. The signal has much parameters. The most appropriate features of speech recognition systems are Mel-frequency Cepstral Coefficients (MFCCs), which can characterize the human auditory about speech well. The speech signal is considered to be short time stationary within 20 to 30ms. In this system, frame length is 20ms, and frame shift is 10ms.

### C. *Acoustic Model*

Acoustic model is a signal processor, which transforms the speech waveform into a sequence of feature vectors. Building the acoustic model includes pre-processing, feature extraction, acoustic model training and test.

The Hidden Markov model (HMM)[4]is applied in this system. HMM can be viewed as a finite state automata which is in a state at any given time. The state is related to initial state *a* and state transition probability $A(a_{ij})$, which is time-independent. Corpora are used to train HMM model.

### D. *Model Training*

Acoustic model training is the kernel of the whole system. Because it affects the performance of the overall system directly.

Co-articulation in the speech makes many variants in different context. It can be abstracted as tri-phone model. In this system, the basis modeling units are context-sensitive phonemes [2]. There are 50 phonemes. Besides, in Chinese, we can use initials and finals as basis modeling units. There are 66 initials and finals[7].

In acoustic model, corpora to be recognized may not match the training corpora. This will get poor performance. To make domain adaption, self-adaption[3]approach is used.

### E. *Dictionary*

In order to output words, dictionary must be build. There are two kinds of dictionary. The first dictionary contains 406 Chinese single syllables. The second dictionary contains 2245 words, which includes frequently-used place names in Beijing and some event words. If the recognition results match the words in the dictionary, it will output corresponding words.

### F. *Recognition Network*

Recognition network contains dictionary, language model or grammar, it defines the candidate set to be output in speech recognition system. The dictionary is to define pronunciations of words in vocabulary. Language model makes the output sentences obey the grammar.

Given sentence set S, and speech signal a, the goal of speech recognition is to find out the corresponding sentence that satisfies following formula.

$$\underset{s\in S}{\operatorname{argmax}}(P|(s|a))$$

According to the Bayesian rule,

$$\underset{i\in S}{\operatorname{argmax}}\big(P\big|(s|a)\big) = \underset{s\in S}{\operatorname{argmax}}\, P(a|s)P(s)/P(a)$$

The prior probability P(a) of speech signal is not related to s, so it can be ignored. P(a|s) is known as relation between sentence and signal, so it is called acoustic model. P(s) is language model that represents sentence probability in the language. The most widely used statistical language model is N-gram.

## IV. STRUCTURED INFORMATION EXTRACTION

Keywords sequences cannot be used by the decision module of the vehicle directly. Besides, they contain a lot of useless information that is not related to traffic condition. It is necessary to convert the keywords into structured information. The whole traffic condition message contains following slots (shown in Table 1).

TABLE I. FULL SLOTS

| Tag name | Description |
|---|---|
| **LOC_NAME** | Site name. For example, “蓟门桥”, “北皋” |
| **ROAD** | Road name. For example, “中关村大街” |
| **SECTION** | Road section name. For example, “苏州桥到四通桥” |
| **DDIREC** | Direction. For example, “北向南” |
| **CHAN** | Channel. For example, “最外侧车道” |
| **CONDIT** | Condition. For example, “车辆行驶缓慢” |

A sentence contains all the above slots, or part of the above slots. A slot contains one or more words. The example is shown in Figure 3.

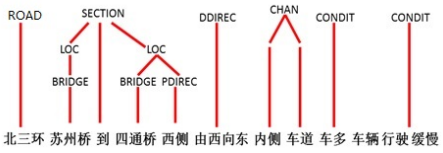


FIGURE III. COMPONENTS OF A SENTENCE

To solve this problem, it should make partial parsing. Chunks are non-overlapping regions of words. A chunk contains one or more words, and can be given a tag. It is a sequence tagging problem. Every word is classified a tag, such as “ROAD-B”, “SECTION-B”, “SECTION-I”, “CONDIT-B”. The prefix of the tag name is the class of the chunk. If the suffix is “B”, it is the beginning of the chunk. If the suffix is “I”, it is not the beginning of the chunk. If the words is not related to the events or place, it can be tagged “O”.

The solution to sequence tagging is Hidden Markov Model (HMM) [4], Maximum Entropy (ME) [5] and Conditional random field (CRF) [6].

As a generative model, the N-gram LM underlying the HMM transition model makes it difficult to use features that are highly correlated (such as words and POS labels) without greatly increasing the number of model parameters. In contrast, MaxEnt gives multiply kinds of feature function. However, it attempts to make decisions locally, without using sequential information. CRF combines the benefits of HMM and Maxent. So we use CRF here.

ROAD: 北三环
LOC_NAME: NULL
SECTION: 苏州桥到四通桥西侧
DDIREC: 由西向东
CHAN: 内侧车道
CONDIT: 车辆行驶缓慢

FIGURE IV. EXAMPLE OF OUTPUT FORMAT

To extract features, dictionary is built which share the same words list as that in keywords output module. Also, the words list contains all of the place name in Beijing. Every word is given a class, such as ROAD, BRIDGE, DIRECTION, and CONDITION. Word and its class, together with the context, are extracted as features. They are trained by CRF model.

After chunk parsing, it fill the slots. The example format is shown in Figure 4.

## V. EXPERIMENTS

In our experiments, every tri-phone model is represented by a 5-states HMM model. The output distribution of each state is represented by diagonal Gaussian Mixture Densities with 16 components. The features here are 39 dimension MFCCs. The training data of acoustic model are 86453 wav files of "863 corpora" and 11460 wav files of traffic broadcast corpora.

To improve the performance, Bi-gram language model is used. The training corpora include 3400 effective sentences of traffic broadcast, 20998 Microblog sentences of traffic domain and 724 sentence from Beijing Traffic Information website. The test corpora contain 7936 effective sentences of traffic broadcast.

### A. *Performance of Acoustic Model*

To evaluate the model, experiments are conducted on the training data using dictionary of single syllable and keywords. Also, to evaluate the performance in language model, Bi-gram is used in each method. The correct rates and accuracy of syllables in each method are shown in Table 2.

TABLE II. PERFORMANCE OF ACOUSTIC MODEL

| **Method** | **Bi-gram** | **Rule based method** |
|---|---|---|
| **Single syllable** | Corr: 51.82% Acc: 47.67% | Corr: 33.94% Acc: 17.01% |
| **Keywords** | Corr: 56.42% Acc: 51.32% | Corr: 37.18% Acc: 22.15% |

### B. *Performance of Keywords Recognition*

To the keywords recognition performance, precision, recall and F-measure are calculated. The data is shown in Table 3.

TABLE III. PERFORMANCE OF KEYWORDS RECOGNITION

| **Precision** | **Recall** | **F-measure** |
|---|---|---|
| 54.88% | 42.95% | 48.19% |

### C. *Performance of Partial Parsing*

Strictly, to evaluate the overall system, just calculate the accuracy of sentences. If all the tags of a sentence are recognized correctly, the sentence is correct. According to the experiment, 806 out of 7936 sentences is correct completely.

Then we can calculate the tagging accuracy of a sentence, and count the number of sentence in accuracy range. The figures are shown in Figure 5.

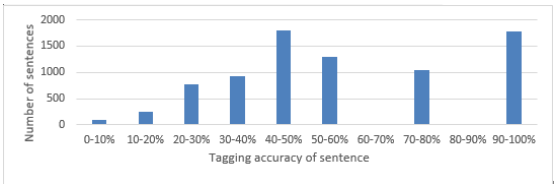


FIGURE V. PERFORMANCEOF PARTIAL PARSING

Though the number of completely correct sentences is small. The tagging accuracy of sentences distribution shows it is not so bad. As slots in the sentence are not all important to the whole tips. The recognition error on unnecessary slots effect less on the path decision.

## VI. DISCUSSION AND CONCLUSIONS

This paper introduces approach of key information extraction from Broadcast in traffic domain, including Speech Recognition and Natural Language Processing.

The system pre-processes the original speech signal, and makes speech enhancement. Then it builds the HMM model based on Tri-phone model in Initials and finals. Lately, it outputs keywords based on n-gram. The dictionary of single syllables differs with that of keywords in performance. Approach based on keywords perform a little better than that based on single syllables.

In language processing module, it uses CRF to make partial parsing. At last, it fills the slots about the chunks. So that it can output structured traffic information. From the experiments, we find that n-gram can improve the performance of the whole system. The results shows the difficulty on speech recognition. Also, the mature products in industry perform poor in a special domain, even with the strong background music.

So we conclude that our results is satisfying. How to build the robust interface between each modules is the key point we should study on.

## ACKNOWLEDGEMENTS

This work is supported by National Nature Science Foundation of China (NSFC): Grant No.91120015 and No.11161140319.

## REFERENCE

[1] Ming Tu, Xiang Xie, Yishan Jiao, "NMF based speech and music separation in monaural speech recordings with sparseness and temporal continuity constraints", ICMT, pp. 548-555, 2013.

[2] R.Schwartz, Y.Chow, O.Kimball, S.Roucos, M.Krasner and J.Makhoul,"Context-dependentModeling for Acoustic-phonetic Recognition of Continuous Speech", ICASSSP,1985,PP.1205-1208

[3] J. McDonough, T Schaaf, A. Waibel. On Maximum Mutual Information Speaker Adapted Training. Acoustics, Speech, and Signal Processing, 2002 Proceedings. IEEE International Conference. 2002:601-604

[4] Rabiner L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE,77(2): 257-286.

[5] Berger A L, Pietra V J D, Pietra S A D. 1996. A maximum entropy approach to natural language processing. Computational linguistics, 22(1): 39-71.

[6] Sutton C, McCallum A. 2006. An introduction to conditional random fields for relational learning. Introduction to statistical relational learning, MIT Press.

[7] Zhang J, Zheng T F, Li J, et al. Improved context-dependent acoustic modeling for continuous Chinese speech recognition[C]. INTERSPEECH. 2001: 1617-1620.