Intention Understanding Based on Multi-source Information Integration for Chinese Mandarin Spoken Commands

Jia Jia, Yongxin Wang, Zhu Ren, Lianhong Cai
Department of Computer Science and Technology, Tsinghua University
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology(TNList)
Beijing 100084, China

***Abstract*— In this paper, we address the problem of Intention understanding for Chinese Mandarin spoken commands. Unlike the previous works, we propose an intention understanding approach including not only the detection of command content, but also the detection of user's affective state. For command content detection, we propose a strategy of keyword combinations analysis using concept restrictions, based on N-best speech recognition results. For affective state detection, we propose a method of multi-source information integration in decision level, using a weighted maximum confidence score to get a high reliability for both acoustic features and speech recognized text. Experimental results show satisfactory effectiveness in command content detection, while combining multi-source information improves the performance of affective state detection.**

Keywords- intention understanding; command content; user's affective state; multi-source information integration

I. INTRODUCTION

Researches on command understanding aim at understanding the intention of speakers from the spoken commands accurately. It is well-known that information delivered by speech is more than the literal meaning. Acoustic information contained in speech is also helpful for understanding. In addition, speaker intention can be highly influenced by context information. Therefore an appropriate understanding of command intention should not only focus on literal meaning alone, it is necessary to take into account other information.

Typical works on speech intention understanding are the researches of spoken language understanding (SLU) [1], which analyzes the transcribed text of a domain-specific spoken speech to obtain its semantic meaning. Existing SLU methods are quite effective in extracting literal meaning from speech. However, most of them neglect information like speaker's affective state or context, which may even lead to a total misunderstanding of the real intention.

For the purpose of understanding command intention more effectively, we make an initial effort in extracting and combining information from multiple sources including recognized text and acoustic features, which is briefly shown in Figure 1.

The main contributions of this paper are the following:

- The command intention is broadened by incorporating affective state with command content;
- Command content is detected by analyzing the combination of keywords recognized from speeches, which can be adapted to different kinds of applications.
- The speaker's affective state is determined by combining the information extracted from acoustic features. These results are described by a 3-dimentional space (VAD, Valence-Activation-Dominance), and calculated through the confidence of each source. The speaker's affective state indicates the reliability of information integration.

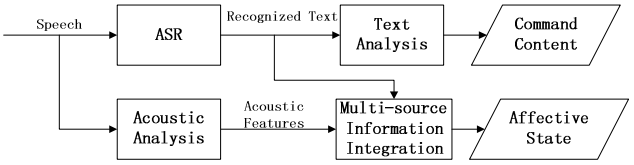

Figure 1: *The proposed intention understanding approach.*

The rest of this paper is organized as follows: in section 2, we describe the intention definitions used in the proposed approach, and the framework of our intention understanding approach. Section 3 describes the algorithms and implementations in details. Section 4 shows some experimental results. Conclusions are discussed in section 5.

II. INTENTION DEFINITION AND APPROACH FRAMEWORK

In order to describe our method more clearly, we take the domain of home automation as our intention understanding scene. Home automation refers to the automation of housework or household activities. User can control the household appliances with speech commands. Therefore, the intention to be detected is defined to a set of household tasks. In additional, user's affective state is considered as a part of intention as well. In other words, the intention information includes not only "what to do" but also "how to do".

Definition 1: **Intention** of a speaker contains two parts: command content, and user's affective state.

Definition 2: **Command content** shows the "to do" task. Command content contains five key concepts: "Device" refers to the device to be manipulated (e.g. a TV); "Attribute" means the attribute to be adjusted (e.g. volume); "Operation" stands for the operation to be taken (e.g. turn up); "State" shows the

current state of the device to be operated on; and "Position" indicates the location of the device to be manipulated.

Definition 3: **Affective state** describes "how to do" the task. Affective state contains one key concept which describes user's feeling when he/she speaks. We model the user's feeling in a 3-dimension VAD emotion space [2]. V (Valence) indicates whether an emotional state is positive or negative, A (Activation) represents the excitation level, and D (Dominance) shows the apparent strength of the speaker. We define 3 levels at each dimension: -1 (negative/low), 0 (neutral/normal), and 1 (positive/high).

Affective state may influence the priority of a command which relates to the tolerance to the length of response time: If the level of D is determined as 1, the priority should be high; if the VAD are all -1, the priority of the corresponding command should be very low.

A. Key Concepts and Keywords

For natural human computer interaction, we allow users to speak in a natural way. However, spontaneous speech always contains repetitions, omitting, reversal, etc [3], which will lead to a severe degradation of the recognition performance. On the other hand, not every part of a sentence is useful for intention understanding, and a few key concepts can effectively represent the meaning of the whole sentence. Therefore key concepts are adopted to represent the intention of speech.

Take the application of home automation as an example. Six key concepts are defined, including device type (cc_device), device position (cc_pos), device attribute (cc_attri), device state (cc_state), available operation (cc_oper) and user feeling (us_feeling). They are a few keywords in each key concept, such as "open" and "close" are of the type "available operation", as shown in Table 1. The command content in our framework is represented by combinations of keywords, regardless of their order of appearance.

Table 1. *Key concept and keywords*

Intention	Key Concept	Keywords Example
Command Content	cc_oper	打开,开(Open)
	cc_device	灯,电灯(Light)
	cc_attri	风力(Wind)
	cc_pos	卧室(Bedroom)
	cc_state	高,大(High)
Affective State	us_feeling	赶快，真着急 (Hasty)

B. Context Information

Besides the above key concept, we also consider using context information to help the intention understanding. Context is represented by the same five key concepts in Table 1. The context information used in this work mainly refers to the scene information, such as the user's position. For example, when we cannot find the position information of a device cc_pos from the input speech, we could take the user position into consideration based on the assumption that user is more likely to manipulate the devices nearby. Context information can also help the user speak in a much more natural manner. For instance, when a user wants to turn on a light and the power of the target light is off, then the user doesn't need to say the whole command "Turn on the light", but "Light" only.

Here gives the framework of the proposed intention understanding approach, which is shown in Figure 2.

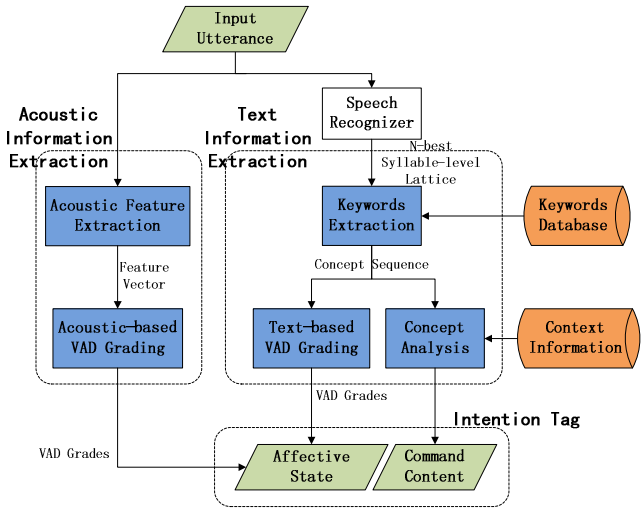

Figure 2: *Framework of the intention understanding using multi-source information.*

For command content detection, we presented a concept analysis algorithm to combine the command-related keywords with context information. At the same time, we proposed a weighted maximum confidence score combining rule for user's affective state detection, using extracted acoustic features as well as emotional keywords.

III. ALGORITHMS AND IMPLEMENTATIONS

A. Keyword Extraction Based on N-Best Lattice

Keywords extraction methods generally use a garbage model to capture non-keywords [4]. Such methods are suitable for real-time applications, but adding new keywords can be very expensive.

To ensure the scalability of keyword database, a LVCSR (Large-Vocabulary Continuous Speech Recognition) system for mandarin is adopted, which is constructed using Microsoft SAPI5.1. The LVCSR system generates an *N*-best lattice using recognized Chinese characters with phonetic transcriptions (pinyin). An *N*-best lattice is in the form of *N* time-aligned recognition results with highest recognition scores. Using *N*-best lattice instead of 1-Best can provide more information for keyword extraction while not causing large searching overhead [5], since the lattice has a standardized structure.

1) Keyword Confidence Measure

In order to calculate the similarity between different candidate syllable strings and the correct phonetic transcription of a keyword, we use sub-syllabic unit. A

Chinese syllable consists of three parts—initial, final and tone. Each part of them is regarded as a sub-syllabic unit.

In the recognition result of N-best lattice, the syllables are organized by the recognizer as segments of words. The syllables in different recognition results are aligned, and place holders would be inserted into appropriate positions in sentences with fewer syllables. In this way, all the recognized sentences would have the same number of syllables. In the process of keyword recognition, each word in the recognition result would be called a candidate element, which would be used to form a keyword. The candidate elements in the lattice (with is formed by the N best result) would be represented by $[e;n_b,n_s]$, where n_b and n_s are the positions of the first and last syllable of the element, while the first syllable in the sentence has position 1. The set of all the candidate elements in the ith recognition result would be noted as E_i. A segment in the ith recognition result would be noted as $[E_i;n_1,n_2]$ ($n_1 \le n_2$), which can be a candidate element or not.

An acoustic score is assigned to a candidate element by the recognizer, which will be noted as $A([e;n_b,n_s])$. We assign this acoustic score to each syllable inside the candidate element, noted as $A^i_n = A([e;n_b,n_s])$, $e \in E_i$, $n_b \le n \le n_s$.

To calculate the acoustic score of a possible keyword, we cumulated the acoustic scores of the candidate elements which form the keyword. If a possible keyword k locates at position n_1 to n_2 in the recognition result, the acoustic score of a keyword can be defined as follows:

$$ACM(k) = \sum_{i=1}^{N} NA([E_i;n_1,n_2]) \cdot \frac{SU(k \| [E_i;n_1,n_2])}{SU(k)} \qquad (1)$$

where $SU(k \| [E_i;n_1,n_2])$ indicates the matched number of sub-syllabic units between the segment $[E_i;n_1,n_2]$ and keyword k, and $SU(k)$ is the total number of sub-syllabic units that keyword k contains. $NA([E_i;n_1,n_2])$ is the normalized acoustic score of segment $[E_i;n_1,n_2]$, which is calculated as

$$NA([E_i;n_1,n_2]) = \sum_{n=n_1}^{n_2} A^i_n \Big/ \sum_{j=1}^{N} \sum_{n=n_1}^{n_2} A^j_n \qquad (2)$$

Given a keyword set $\{k_1, k_2, \cdots, k_N\}$ and a command content set $\{p_1, p_2, \cdots, p_M\}$, we use the frequency that two keywords appear in the same command to evaluate the semantic co-occurrence probability between these keywords. The definition of co-occurrence probability of keyword k_i to keywords k_j is

$$SS(k_j|k_i) = \frac{count(k_j,k_i)}{\sum_{n=1}^{N} count(k_n,k_i)} \quad (1 \le i,j \le N) \qquad (3)$$

where $count(k_j,k_i)$ is the number of co-occurrences of keywords k_j and k_i in the command content set.

Considering that a keyword is much more likely to construct one command content with adjacent keywords, we compute the semantic score of a keyword by summing up all the co-occurrence probabilities between this keyword and its adjacent keywords.

$$SS([k;n_b,n_e]) = \sum_{\substack{\text{all } k' \text{ whose} \\ n'_e = n_b - 1 \text{ or} \\ n'_b = n_e + 1}} SS([k';n'_b,n'_e] | [k;n_b,n_e]) \qquad (4)$$

where $[k;n_b,n_e]$ is a possible keyword k start at n_b and ends at n_e; $SS([k;n_b,n_e])$ is the semantic score of keyword k.

2) Keyword Extraction Algorithm

In the process of keyword extraction, the candidate elements from the recognition results are considered as possible keywords, as the candidate elements are all considered words in the recognition result. However, there are segmentation errors in the recognition result, and that may cause some keywords not to be found. To solve this problem, we further improved the algorithm. For segments where no successful match is found when only considering candidate elements, a rescan process is carried out disregarding the candidate element boundaries using the Backward Maximum Matching (BMM) segmentation algorithm, to find the missed keywords when only considering the candidate elements.

After the keywords are identified, their acoustic scores will be computed. If there are several keywords with the same key concept at the same position, they will be merged to one keyword, whose acoustic score is the accumulation of the keywords merged. The semantic score of each keyword is also calculated at the same time. The keyword with both the lowest acoustic score and semantic score at one position will be dropped.

B. Command Content Analysis using Key Concept Relation

Common command control systems restrict the user's expression on the order of keywords' appearance. To solve this problem, a command content analysis strategy is designed to detect command content despite different combinations and orders of the input keyword.

1) Restriction of Key Concept Relation

Key concepts in a meaningful command should comply with the concept restrictions, which are determined by the scene and command pattern. Restrictions to the five key concepts related to command content are described in details as follows.

When the key concepts are ordered as: "cc_pos", "cc_device", "cc_attri", "cc_state" and "cc_oper", the restrictions only exist between two adjacent concepts. Some specific "cc_device" can only be related to some specific "cc_pos", e.g. TV can only be found in bedroom and living room; "cc_attri" is subordinate to "cc_device"; e.g. the relationship between volume and TV; "cc_state" is subordinate to "cc_attri", e.g. the level of volume; "cc_oper" is restricted by "cc_state", e.g. a light can be turned on only when it is off.

As shown above, a meaningful command should contain five key concepts, whose combination should agree with the provided restrictions.

2) Command Content Analysis

Command content is analyzed by searching for one or more meaningful combinations of input key concepts, whose matching positions are close enough.

The analysis process can be divided into two stages. In the first stage, key concepts in the input sequence are examined to find out all other concepts that can be inferred from them. Each inferred key concept is added to the sequence and keep a copy of the position value of its source concept.

In the second stage, meaningful combinations of key concepts with the closest matching positions are picked out.

Context information is taken into consideration at this stage. Context is represented by a set of five key concepts. Once an incomplete combination is extracted, proper key concept in the context concept set can be filled in.

C. Affective State Detection using VAD Grading and Information Integration

1) Acoustic-based VAD Grading

Acoustic VAD grading is adopted from general approach of speech emotion classification [8].

Applying appropriate acoustic features is a key to emotion classification. Currently, the most widely investigated features are prosodic features, voice quality features and spectral features. Studies have proved that, emotion information in speeches is mainly reflected in the variation of prosodic features [9]. Prosodic features are the statistical and temporal features related to fundamental frequency, duration and energy.

Prosodic features including mean value of short-time energy, average syllable duration, mean value of fundamental frequency (F0), F0 maximum and range value are exploited in this work. All the features are extracted using algorithms provided by Praat [10].

Support vector machine (SVM) has been proven to achieve satisfactory performance for classification task, while requiring a small amount of training data [11]. For each dimension of VAD, a SVM classifier for three-class classification (negative/low, neutral/normal, and positive/high) is constructed. All three classifiers are implemented by libsvm [12]. A <V, A, D> level set can be obtained from the classifiers. Each value in the set is labeled with its confidence score from the SVM algorithm, which is a real value between 0 and 1.

2) Text-based VAD Grading

To determine user's state, feeling related text keyword and VAD levels extracted from the VAD Grading module are merged. All the text keywords of type "feeling" are labeled with VAD values [2]. For example, "angry" is labeled with <1, 1, 1>, "sad" is labeled with <-1, -1, -1>.

If "feeling" keywords are detected, the V value of the speech from text will be determined by voting:

$$\begin{aligned} C_{V,k} &= \frac{1}{n}\sum_{i=1}^{n} CW_{V=k,i}, \quad k=-1,0,1 \\ V_T &= \arg\max_{k=-1,0,1}(C_{V,k}) \\ CV_T &= \max(C_{V,k}) \end{aligned} \tag{5}$$

where $CW_{V=k,i}$ means the confidence score (similarity score derived from keywords extraction) of the ith word whose V value is k. $C_{V,k}$ is the mean value of $CW_{V=k,i}$. T means "from text". The maximum value of $C_{V,k}$ is considered as the confidence score (CV_T) of V for the whole speech and the corresponding k is taken as the value of V for the whole speech (V_T). A_T and D_T can be determined in the same way.

3) Multi-source Information Integration

Multi-source information integration can be carried out at feature-level or decision-level. Since the acoustic features and text features are different in content and form, we execute integration at decision level to obtain the key concept of user state (us_feeling).

The confidence of each information source is not carefully studied in most of the existing decision-level information integration methods as [8], which may affect the reliability of integration result. To solve this problem, we proposed an integration method making full use of the single-source confidences.

Define V_S, A_S, and D_S as the V, A, D value extracted from acoustic information, their confidence scores are CV_S, CA_S, and CD_S (derived from Acoustic-based VAD Grading), respectively. Then the final V value is determined by a weighted maximum confidence score combining rule as follows:

$$V = \begin{cases} V_S & \mu_V CV_T \le (1-\mu_V)CV_S \\ V_T & \mu_V CV_T > (1-\mu_V)CV_S \end{cases} \tag{6}$$

where μ_V is determined by the maximum grading accuracy rate when applied to training corpus. T and S refer to "from text" and "from acoustic speech" respectively. A and D are determined in the same way.

IV. Experimental Results

We conducted two objective experiments to evaluate the performance of command content analysis and affective state detection respectively.

A. Experiment on Command Content Analysis

The command content mentioned in this experiment is related to the home automation scene. The testing Chinese corpus was recorded by five people (2 female and 3 male). The corpus has 1000 speeches containing 5260 keywords and lasting for almost 70 minutes.

We choose different keyword extraction methods for comparison. The baseline system chosen to compare with our methods is the keyword recognition algorithm based on string matching, using the log-likelihood ratio score as the confidence measure for keyword spotting. We also compare the performance of keyword extraction without (original system) and using (improved system) the rescan process. All

these three methods use the same command content analysis method mentioned in III.B.

The Detection Rate (DR) and False Alarm Rate (FAR) are selected to evaluate the performance of the command content analysis methods. They are defined as follows:

$$DR = \frac{correctly\ spotted\ keywords}{total\ keywords\ in\ the\ speech} \times 100\% \quad (7)$$

$$FAR = \frac{incorrectly\ spotted\ keywords}{correctly\ and\ incorrectly\ spotted\ keywords} \times 100\%$$

Experimental results are shown in Figure 3.

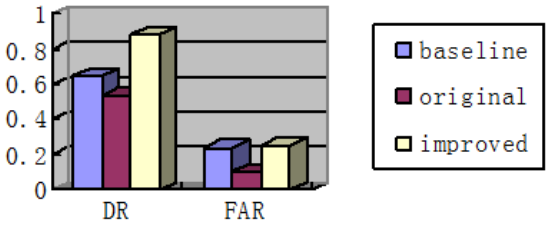

Figure 3: *Performance Comparison of Command Content Analysis.*

As we can find from Figure 3, the original keyword extraction algorithm had the lowest DR, because the baseline system aimed at matching one candidate speech at a time. The processing primitive of the original keyword extraction algorithm is element so that it was easily affected by the segmentation error occurred in the front-end speech recognition system. But fortunately, we can find that the DR of the improved keyword extraction algorithm was much better than the baseline system, and their FAR are comparable. These results indicate the effectiveness of our command content analysis method.

B. Experiment on Affective State Detection

For user's affective state detection, a corpus contains 4 subsets with different affective states (angry, happy, sad and surprise) was applied. Each subset has 220 speeches and labeled with a VAD tag manually. The contents of these speeches are not limited to speech commands. 20 command speeches of each subset labeled as "hasty" are also recorded, since hastiness is a helpful emotion in the speech command scene. Among each subset, we choose 200 speeches for training and the remaining 20 for testing.

Classification accuracy for V (P_V) is defined as the number of speeches assigned with a correct V value versus the number of all tests. P_D and P_A are defined in a similar manner. When using acoustic information only, we get a classification accuracy of 87.5% for V, 88.75% for both A and D.

With the help of recognized text, which means we integrate text information with acoustic information for user's affective state detection, both the classification accuracy for V and A has an improvement of 5% (the value of μ was set to 0.45), but the classification accuracy for D has no obvious improvement. We think there are three reasons: 1) the affective state is more likely to be contained in acoustic information; 2) most keywords about us_feeling are adjective words describing the user's positive or negative state, so the text information has a greater help in V; 3) our corpus has "hasty" training set, the keyword such as "hurry up (赶快、快点)" improve the performance of A. And these results demonstrate the benefits of using multi-source Information Integration for user's affective state detection.

V. CONCLUSION

In this paper, we address the problem of Intention understanding for Chinese Mandarin spoken commands. Unlike the previous works, we propose an intention understanding approach including not only the detection of command content, but also the detection of user's affective state. For command content detection, we propose a strategy of keyword combinations analysis using concept restrictions, based on N-best speech recognition results. For affective state detection, we propose a method of multi-source information integration in decision level, using a weighted maximum confidence score to get a high reliability for both acoustic features and speech recognized text. Experimental results have shown satisfactory effectiveness in command content detection, while combining multi-source information utilizing single-source confidences enhances the performance of affective state determination.

In the future work, the confidence measure of text information and acoustic information will be further studied to improve the performance of semantic information integration.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (61003094, 90820304，90920302). And this work is supported by Tsinghua - Tencent Joint Laboratory for Internet Innovation Technology, and funded by Tsinghua National Laboratory for Information Science and Technology（TNList）Cross-discipline Foundation.

REFERENCES

[1] Y.-Y. Wang, L. Deng and A. Acero, "Spoken Language Understanding," IEEE Signal Processing Magazine, vol. 22, issue 5, pp.16-31, Sep. 2005.
[2] R.Kehrein, "The Prosody of Authentic Emotions," Proc. Speech Prosody Conf., pp.423-426, 2002
[3] CQ. Zong, H. Wu, "Analysis of Spoken Dialog Corpus in Restricted Domain," Proc. JSCL-99
[4] Mporas, I. et al. "Automatic Speech Recognition System for Home Appliances Control" Proc. of the 13th Pan-Hellenic Conference on Informatics 2009, 114–117.
[5] Timo B, Okko B. et al. "Evaluating the Potential Utility of ASR N-Best Lists for Incremental Spoken Dialogue Systems," Proc. Of INTERSPEECH 2009, 1031-1034.
[6] G. Navapro, "A Guided Tour to Approximate String Matching," ACM Computing Surveys, vol.33, pp.31-88, 2001
[7] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification. John Wiley & Sons, 2001.
[8] CM. Lee and S. Narayanan. "Towards Detection Emotions in Spoken Dialogs," IEEE Transaction on Speech and Audio Processing. 2005. Vol.13, no. 2. pp: 293:302.
[9] J.Tao, T.Tan and R.W.Picard, "Feature Importance Analysis for Emotional Speech Classification," ACII 2005, LNCS 3784, pp. 449-457, 2005
[10] P.Boersma and D.Weenink, Praat: doing phonetics by computer [Computer program]. Version 5.1.34, retrieved 31 May 2010 from http://www.praat.org/
[11] H. Hu, MX. Xu and W. Wu, "GMM Supervector based SVM with spectral features for speech emotion recognition," Proc of ICASSP, 2007, 413-416.
[12] C-C.Chang and C-J.Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm