Applied Mathematics & Information Sciences
An International Journal

Speech Recognition with Word Fragment Detection Using Prosody Features for Spontaneous Speech

Jui-Feng Yeh [1] and Ming-Chi Yen [2]
Dept. of Computer Science and Information Engineering, National Chiayi University, Chiayi City, Taiwan (R.O.C.)
Corresponding author: Jui-Feng Yeh, Email: ralph@ncyu.edu.tw[1]

Received June 22, 2010; Revised March 21, 2011; Accepted 11 June 2011

Abstract: This investment proposed a novel approach for word fragment detection with prosody features for spontaneous speech recognition. Incomplete pronunciation of word result in ill-form fragment in word-building that causes the performance of language model in speech recognition is dramatically decreased. Instead of lexical word, prosody word is used to be building block for spontaneous speech processing recently. Prosody features are further extracted from prosody word and fed into the decision tree to judge the prosody word is complete word or word fragment. There are three categories feature sets are employed here: pitch related, intensity related, and duration related features are included. For evaluating the proposed method, the Hidden Markov models (HMMs) based speech recognition core was developed to be the baseline. The proposed method is integrated into the baseline to provide the word fragment detection capability and enrich the performance of spontaneous speech recognition. According to the experimental results, the performance of proposed method outperforms traditional speech recognition especially in insertion and deletion error. This shows that the word fragment detection can obtain the improvement for spontaneous speech recognition.

Keywords: Word Fragment, spontaneous speech recognition, prosodic feature, decision tree.

1. Introduction

Conversational speech or spontaneous speech is increasing essential for human daily life. To be one of the most important human machine interfaces, speech recognition technologies play key roles for intelligent systems' perception. In practice applications, spontaneous speech recognition is the trend of information technology applied to real word in the near future. However, there are many factors affect the spontaneous speech recognition results very much. In conversational speech, speech act identification is very essential for spoken language understanding. Yeh et al. proposed an ontology-based speech act identification base on partial pattern tree [1]. To understand such problem, the composition and characteristics of spontaneous speech are first analyzed. Considering of the composition about spontaneous speech, it can be divided into several steps: Think about themes of the dialogues and then find the relevant parts of memory. Constitutes a statement and speak out statements [2, 3]. According to the composition procedures, speakers repeat the steps to generate utterance for a continuous conversational dialogue. If errors occurred in any step, disfluencies will appear. Actually, word fragments usually co-occur with disfluency according to the observation of phenomenon about utterances. Unfortunately, word fragment reduces the performance of speech recognition significantly. Many insertions and deletions will be produced by conventional approach, thereby accuracy is dropped. The word fragment detection is able to reduce these errors to enhance recognition performance.

The labeling of disfluence in the spontaneous speech is very helpful for the semantic understanding according to the concluding of related works in [4] and [5]. Bear et al. used the pattern matching, parsing and acoustic models to detect the disfluence in speech [6]. Nakatani and Hirschberg [7] created the repair interval model to predict the repair in speech. From the observations about these works, it is found that

more than half disfluency speech contains word fragment. It means that the words fragments occur frequently in disfluency speech also are good indicators. About the related works about word fragment, Dan Jurafsky [8] compared the support vector machine (SVM) and decision tree with lexicon feature, prosody feature and voice quality to detect the word fragment of Mandarin in phone conversation. Liu [9] pointed out that the word fragments correlation with spontaneous speech, using the prosody features. Yeh and Wu used the prosodic features to detect the potential interrupt points of the disfluency in spontaneous speech [10]. The features in their detection rules were energy, pitch, duration and voice quality related. Additionally, Liu et al. [11] compared the hidden Markov model (HMM), Maximum entropy (ME) and conditional random field (CRF), using different structures with different features to detect the interrupt point in disfluency speech. Besides, Lee et al. presented new set features: duration-related and pitch-related [12] for disfluency detection of spontaneous speech. They detected disfluency interruption point (IP) for spontaneous Mandarin speech. Lee et al. in [13] presented further works, Lee used latent prosody modeling and incorporated decision tree into maximum entropy model (DT-EM) was developed. In prosody research, Tseng et al. define the prosody boundary to tagging prosody word (PW) and prosody phase (PP) boundary [14]. In 2008, Tseng et al. propose the hierarchical phase group (HPG), prosody related features should consider the whole prosody structure of utterances [15]. Briefly, we can understand the relevancies of prosody features and the spontaneous speech.

In this paper, we proposed a novel detection algorithm based on the decision tree using prosody features. Instead of lexical word, the prosody words are first segmented from the spoken utterance. The prosodic feature set representing prosody word is composed of pitch related features, intensity related features, and duration related features. The decision tree based word fragment detection framework is used to judge the prosody word is word fragment or not. According to the feedback of the word fragment detection, the acoustic models and language model are both reprocess to generating the word sequence results.

The other sections of this paper organized as follows. In section 2, we presented the system architecture according to the function module. In section 3, we introduced the prosody features analysis about the conversational speech to detect the word fragment. The experimental results are shown in section 4. Finally is the conclusion, we summarize the experimental results and discuss the future work in the last section.

2. System Architecture

The proposed system architecture is divided into two phases: training phase and test phase as show in Figure 1. For access the improvement of proposed method, a hidden Markov models (HMMs) based speech recognition core and tri-gram based language model are both developed to be the baseline system. Since the word fragment is very sensitive for speech recognition result, the proposed method aims at detecting the word fragment according to the characters in speech signals. For more clear illustration, training phase and test phase are described in section 2.1 and 2.2 respectively.

2.1 Training phase

The prosody word segmentations of utterances in the corpus are first decided according to Tseng [15]. By the observation of word fragment phenomena of spontaneous speech, prosody features of the prosody word are further extracted to be the feature for detecting word fragment. For training the models, human labeling for complete word or word fragment is tagged in advance. We labeled the word fragment in corpus, and then we extracted the prosody features by frame [16]. For the observed part of word fragment, we also analyzed the phenomena of before and after word fragment, and added these observations to the prosody feature set, in order to facilitate the establishment of decision tree.

2.2 Test phase

In Test phase, speech signal parameters extraction, and then process the speech recognition. In the result, we can get the intervals of each syllable. Meanwhile, the speech signal were also extracted prosody parameters, used the syllable intervals get on the previous step. The syllable intervals provided the duration information, supported decision tree to determine the word fragment exists or not. The original speech

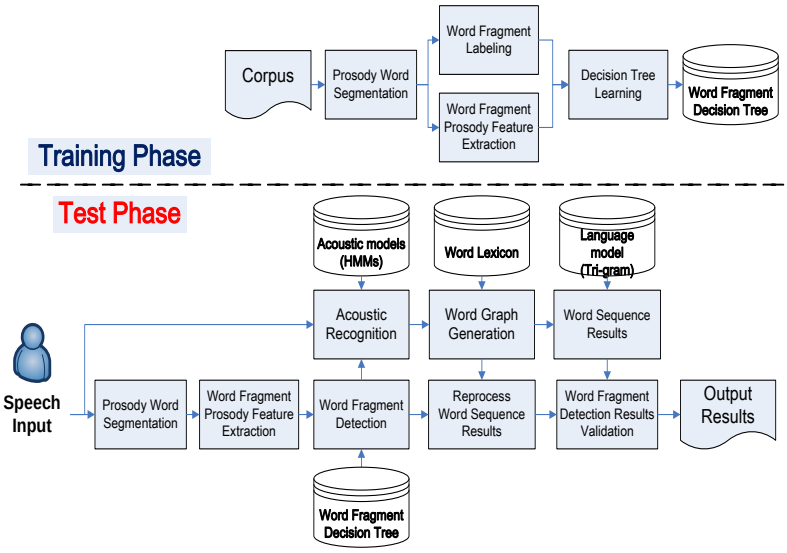

Figure. 1: Schematic diagram of the proposed system architecture

recognition results will as input for language model for generating the word sequence. We use the word fragment detection results regenerate new word sequence, compared new word sequence and the original word sequence to verify the word fragments and correct results.

3. Prosody Feature Analysis

In speech related research, features usually play essential roles for signal processing. For obtaining the near optimal performance, data observations and feature designing are both necessary for word fragment detecting. In this section, the corpus used for evaluation and feature definitions are illustrated as follows.

3.1 Corpus

The Mandarin Conversational Dialogue Corpus (MCDC) is modern spontaneous speech collection. This corpus gathered from 2000 to 2001 by the Institute of Linguistics of Academia Sinica in Taipei, Taiwan. This corpus total length is 27 hours, include 30 conversations. The public released eight dialogues with 15 volunteers (nine female and seven male speakers). The length of recording speech in MCDC is eight hours. The detail information is shown as Table 1. The annotation is labeled according to the definition of Prosody word (PW) in HPG [17] that segmented the utterances into prosody words (PWs) by its spontaneous speech phenomena are proceeded. Each prosody word provided the prosody features including pitch related features, intensity related features and duration related features. We total annotated 171 word fragments wave segments form eight dialogues. The detail descriptions show as Table 2.

Table 1: The corpus details of MCDC

MCDC group	File numbers	Total byte	Total time
001	874	172,791,084	01:29:57
003	1,108	141,261,248	01:13:31
005	910	178,487,666	01:32:55
009	665	165,912,696	01:26:23
010	530	134,426,352	01:10:00
025	671	129,049,748	01:07:11
026	864	133,724,288	01:09:37

Table 2: The word fragment segments of MCDC.

MCDC group	File numbers	Total byte	Total time
001	29	11,956,536	00:06:13
003	19	7,659,280	00:04:00
005	34	35,281,872	00:18:22
009	16	42,471,744	00:22:07
010	16	26,281,728	00:13:41
025	29	19,592,184	00:10:12
026	28	5,481,440	00:02:51

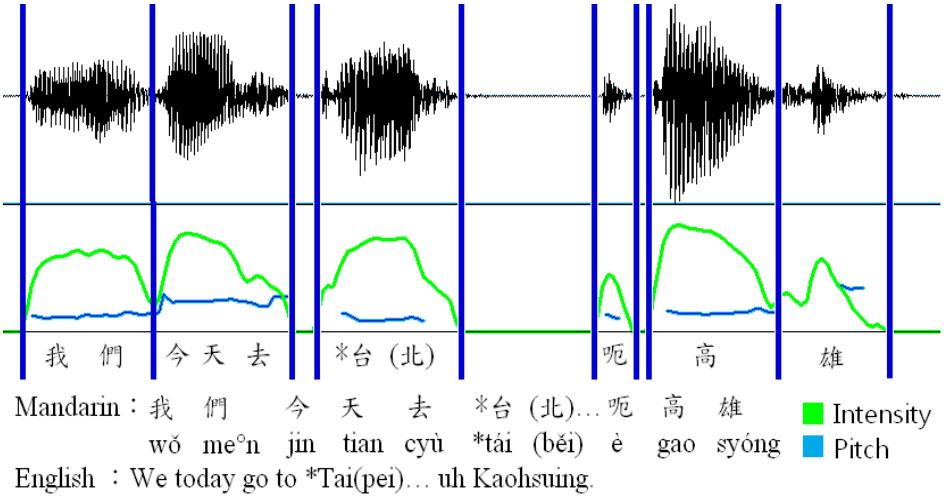

Figure. 2. Repair form word fragment

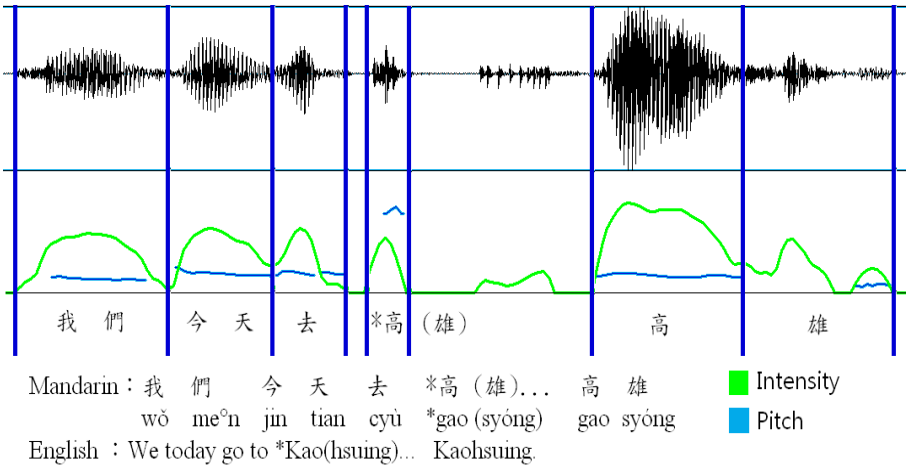

Figure. 3. Restart form word fragment

3.2. The features extracted from word fragment

According to the annotation of word fragment, a set of elementary prosody features are listed as shown in Table 3. The elementary prosody feature set consists of three categories features respectively, from pitch, intensity duration. Considering of the characteristics of contextual co-articulation based on the observation and also considered the feature before and after the word fragment appeared. Similar to N-gram in language models, the sequential feature patterns formed from the preceding prosody word and target word or target words and following prosody word if the target word denotes the prosody word that is decided to be complete words or word fragment. Similar to the disfluency, there are two main categories for word fragment: repair and repetition. Figure 2 represented the repair of word fragment. We can find the word fragment "台(Tai-)" is the subword of "台北(Taipei)" and "台北 (Taipei)" is repaired by "高雄(Kaohsiung)." Figure 3 represented the instant of repetition word fragment. It results from the un-ready or mispronouncing. The first "高(Kao)" means the failure in pronounce for "高雄(Kaohsiung)" in the Figure 3. According to these two mainly word fragment categories' structure, local information similar to N-gram is very essential for detecting word fragment, we call it as sequential feature patterns. Herein, the three main categories of elementary prosody features and sequential feature pattern are described in the follows sections.

3.2.1. Pitch Related Features

Pitch information is very important for speech recognition, especially in pitch trajectory. Two categories of pitch related features are pitch statistical features and pitch trajectory features in this investment. Pitch statistical features includes the minimum value, maximum value and average of pitch in the prosody word. Considering of the pitch trajectory, pitch slope in duration, previous duration pitch slope, and next duration pitch slope are used here. For capture simple information about the trajectory trend, we classify the pitch trajectories into two categories: rise horizon and down horizon. In spontaneous speech, pitch reset is one of the most significantly phenomenon for prosody features especially in prosody word boundary. According to the observation, the characteristics of word fragment are frequently cooccur with pitch reset. Therefore, the related features of the contextual units including the previous and next prosody words are considered in this investment. The pitch trend is calculated using linear regression as shown in equation (3.1).

$$P_t(t) = \alpha_i + \beta_i t, \tag{3.1}$$

where *Pi* (t) denotes *i*-th pitch fragment in time t, βi means the average slope of *i*-th pitch fragment. βi is further formulated as the equation (3.2). denotes the average of time, it represented the middle value of timeline in equation (3.3). The average pitch of *i*-th pitch fragment is estimated as function (3.4). *n* is the number of the sample within the corresponding pitch fragment.

$$\beta_i = \frac{\sum_{t=b_i}^{e_i}(t-\bar{t})(P_i(t)-\bar{P}_i)}{\sum_{t=b_i}^{e_i}(t-\bar{t})^2},\ t \in [b_i, e_i] \tag{3.2}$$

$$\bar{t} = \frac{1}{2}(e_i - b_i) \tag{3.3}$$

$$\bar{P}_i = \frac{1}{n}\sum_{t=b_i}^{e_i} P_i(t) \tag{3.4}$$

3.2.2. Intensity Related Features

Like the energy feature, intensity related feature plays an important role in word fragment detecting. The frame based intensity related features are calculated here; intensity features mainly focus on the variations of pitch trends between two pitch fragments. Generally, intensity usually first rises and then down within the pitch segments. The values of intensity will be extremely fluctuation when the word fragment occurs. Combined with the characters of pitch segments, the intensity short time extremely fluctuation will point out the word fragments. Similarly, the intensity trend and slops are calculated like those of pith related features in equation (1) to (4).

3.2.3. Duration Related Features

Duration indicates the length of each syllable; it can be used to distinguish the prosodic word is word fragment or not. The maximum duration of syllable, maximum duration of syllable average duration of syllables are used to be the features to detecting the word fragment.

Table 3: The list of prosody features.

Feature	Category	Description
P1	Pitch related	Pitch slope in duration.
P2	Pitch related	Maximum pitch.
P3	Pitch related	Minimum pitch.
P4	Pitch related	Average Pitch
P1-1	Pitch related	Previous duration pitch slope.
P1-2	Pitch related	Next duration pitch slope.
E1	Intensity related	Intensity slope in duration.
E2	Intensity related	Maximum intensity.
E4	Intensity related	Minimum intensity.
E5	Intensity related	Average intensity
E1-1	Intensity related	Previous duration intensity slope
E1-2	Intensity related	Next duration intensity slope
D1	Duration related	Duration of observe syllable
D2	Duration related	Maximum duration of syllable
D2	Duration related	Minimum duration of syllable
D4	Duration related	Average duration of syllables

Table 4: Result without word fragment detection

Top 1~5	Acc.	Del.	Sub.	Ins.
MCDC (Top 1)	42.06%	7.79%	27.87%	22.30%
MCDC (Top 3)	43.90%	7.73%	27.48%	20.89%
MCDC (Top 5)	45.20%	7.68%	26.89%	20.41%

Table 5: Result with word fragment detection

Top 1~5	Acc.	Del.	Sub.	Ins.
MCDC (Top 1)	45.26%	8.09%	22.61%	20.89%
MCDC (Top 3)	46.63%	8.663%	22.23%	19.04%
MCDC (Top 5)	48.59%	8.93%	21.89%	18.79%

4. Experimental Result

For evaluating the proposed method, Mandarin Conversional Dialogue Corpus (MCDC) is used as test corpus. There are 100 utterances with word fragment and without word fragment separately are used to access the proposed method. The recognition results of the base line system developed by Hidden Markov models (HMMs) without word fragment detection are shown in Table 4. According to the result, we can find the accuracy of speech recognition is about 42 percentages. There is about 27.87 percentages substation error in speech recognition. Figure 4 shows the precision and recall rates of proposed method to detecting the word fragment. In this experiment, the building block is prosody word. That is to say, we judge the prosody word is word fragment or not after the prosody word segmentation. The best result is 55.83% precision and 66.33% recall; we used this result to produce new word sequence, the recognition result with word fragment detection show as Table 5.

For spontaneous speech recognition, the proposed word fragment detection can reduce the substitution and insertion error significantly.

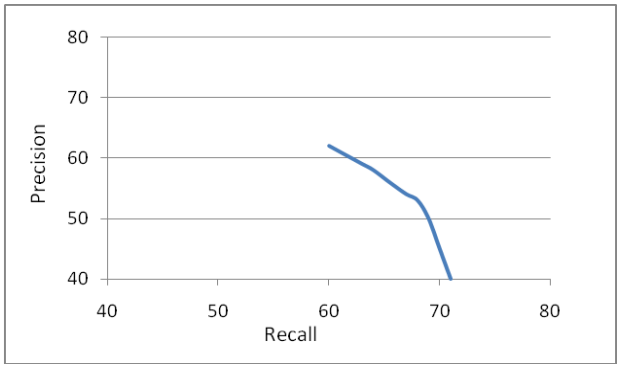

Figure 4: Precision and recall rates of the proposed word fragment detection models

For the new word sequence results, we reduced the substitution and insertion error obviously; the main reason is the word fragment decision tree reduced the number of unrecognized part from the word sequence. The decision tree based word fragment detection using the prosodic features those are extracted from prosody word for spontaneous speech recognition. There are three categories prosodic features are used here, pitch, intensity and duration related features are included. According to the experimental results, we can find the performance is improved significant by integrating the word fragment module into speech recognition core. This shows that the prosody features of prosody word are useful for sub word detecting, and word fragment detection is useful for spontaneous speech recognition. To analysis the experimental results, we can find the improvement mainly come from the correction of substitution and insertion error. However, the deletion rate rises up simultaneously.

5. Conclusion

In this investment, the word fragment detection by decision tree with prosody features is proposed for spontaneous speech recognition. Herein, the pitch, intensity, and duration related features are integrated in a decision tree. Besides the value of the prosody, the trajectory of prosody is considered as the feature for word fragment detecting. According to the experimental results, we can find that the word fragment detecting can significantly improve the performance in spontaneous speech recognition. Decision tree with prosodic features proposed in this paper is effective and efficient for word fragment detection. To achieve more reliable detection result, articulation and contextual information should be considered in the future work. Additionally, the integrating of language model should provide significant improvement for word fragment detection for spontaneous speech recognition.

Acknowledgements

The authors thank the anonymous reviewers for their constructive feedback and helpful suggestions. This work has been supported by National Science Council, Taiwan, under contract NSC 96-2221-E-415-010-MY3.

References

[1] Yeh, J.-F., Wu, C.H., & Chen, M.J. Ontology-based Speech Act Identification in a Bilingual Dialog System Using Partial Pattern Trees. Journal of the American Society for Information Science and Technology, 59(5), (2008), 684-694.

[2] S. L. Yang. Stuttering Research and Treatment Around the World Taiwan, The ASHA LEADER: http://www.asha.org/Publications/leader/2005/051018/f051018a6.htm, accessed on 18 Oct (2011).

[3] S. L. Yang, The Disfluency Loci in Relation to Grammatical Classes:A Psycholingustic Perspective, *Speech-Language-Hearing Association of the Republic of China*, Vol. 16, (2001), 1-19.

[4] E. E. Shriberg, Preliminaries to Theory of Speech Disfluencies, PhD. thesis, University of California at Berkeley, (1994).

[5] Y. Liu, Structural Event Detection for Rich Transcription of Speech, PhD thesis, West Lafayette, Indiana: Purdue University, (2004).

[6] J. Bear, J. Dowding, and E. E. Shriberg, Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog," *Proc. ACL*, (1992), 56-63.

[7] C. Nakatani and J. Hirschberg, A corpus-based study of repair cues in spontaneous speech, *JASA*, (1994), 1603—1616.

[8] C. T. Chu, Y. H. Sung, Y. Zhao, and D. Jurafsky, Detection of word fragments in Mandarin telephone conversation, in *Proceedings of Interspeech 200*6 (2006).

[9] Y. Liu, Word fragment identification using acoustic-prosodic features in conversational speech" in *Proc. HLT Student Workshop*, (2003), 37–42.

[10] J.-F. Yeh and C.-H.Wu, Edit Disfluencies Detection and Correction Using a Cleanup Language Model and an Alignment Model, IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 5, (2006), 1574–1583.

[11] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, Enriching speech recognition with automatic detection of sentence boundaries and disfluencies, *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 5, (2006), 1524–1538.

[12] C. K. Lin, S. C. Tseng and L. S. Lee, Important and New Features with Analysis for Disfluency Interruption Point (IP) Detection in Spontaneous Mandarin Speech, *Proceedings of Disfluency in Spontaneous Speech Workshop*, (2005).

[13] C. K. Lin, and L. S. Lee, Improved Features and Models for Detecting Edit Disfluencies in Transcribing Spontaneous Mandarin Speech, *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 7, (2009), 1263-1278.

[14] C. C. Tseng, *Sinica COSPRO—Corpus and Tools for Mandarin Fluent Speech Prosody, in Computer Processing of Asian Spoken Languages* edited by Shuichi Itahashi, Chiu-yu Tseng, Consideration Books, c/o The Americas Group, 184-188, U.S.A., (2010).

[15] C. Y. Tseng, Corpus Phonetic Investigations of Discourse Prosody and Higher Level Information, *Language and Linguistics*, Vol. 9.3,(2008), 659-719.

[16] P. Boersma and D. Weenink, Praat: doing phonetics by computer (Version 4.5.16), Retrieved Feb 22, (2010), from http://www.praat.org/.

[17] N. Caritey, L. Gaspari, B. Legeard, and F. Peureux, Specification-based testing- Application on algorithms of Metro and RER tickets (confidential). Technical Report TR-03/01, LIFC-University of Franche-Comt´e and Schlumberger Besanc, on, (2001).

Jui-Feng Yeh received the B.S. degree in computer science from National Chiao Tung University, Hsinchu, Taiwan, R.O.C, in 1993. He obtained M.S. and ph. D. degrees in computer science and information engineering from National Cheng Kung University, Tainan, Taiwan, in 1995 and 2006 respectively. He was a Research and Development Engineer and Product manager with the Winbond Electronics Corp. and Advance Multimedia Internet Technology Inc. He is currently an associate professor in computer science and information engineering, National Chiayi University. His research interests include speech signal processing, natural language processing, and knowledge engineering. Dr. Yeh is a member of The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). Email: ralph@mail.ncyu.edu.tw

Ming-Chi Yen is currently pursuing the B.S degree in computer science and information engineering from National Chiayi University, Chiayi city, Taiwan (R.O.C). His research interests include spoken language processing and speech signal analysis. Email: ymchiqq@gmail.com