# Using Keyword Spotting and Utterance Verification to a Prank Call Rejection System

Chun-Jen Lee, Eng-Fong Huang, and Jung-Kuei Chen
李俊仁　　黃英峰　　陳榮貴

Applied Research Laboratory, Telecommunication Laboratories,
Chunghwa Telecom Co., Ltd., Taiwan, ROC
中華電信研究所 應用科技研究室

Email: {cjlee, engfong, jkchen}@ms.chttl.com.tw

## Abstract

*In this paper, a single keyword spotting and verification prototype system aiming at rejecting prank calls is reported. The system issues an announcement in Mandarin which instructs International Operator Direct Connection (IODC) customers to speak a keyword in Mandarin. If the system recognizes the keyword, then it switches the line to a telephone operator. If not, the call is assumed to be a prank call and the line is cut off. The underlying algorithm of this current system consists of a keyword spotter, to extract a single keyword, and a rejector, to verify whether a valid keyword or not, in each spontaneous speech utterance. The experimental results demonstrate that 97.1% of prank calls were rejected while only 2.4% of customer calls were rejected. The field-trial system was developed and has been in operation at the Chunghwa Telecom International Business Group since March 1998.*

***Keyword:*** ***keyword spotting*** ***utterance verification***

## 1. Introduction

Chunghwa Telecom *IODC*, a home country direct service, enables Taiwanese travelers to place a call to a Taiwanese telephone operator directly from overseas. It occupies a large portion of traffic of the overseas incoming calls to Taiwan due to its ease of use. However, over 80% of the incoming calls from some countries are prank calls that seriously compromise the quality of services. To address this problem, a prank call rejection system has been developed which is capable of automatic detecting and rejecting prank calls without connecting to a telephone operator.

During recent years, keyword spotting (Rohlicek 1989, Rose 1990, Wilpon 1990) and utterance verification (Rahim 1995, Kawahara 1997) technologies have become popular methods for domain specific speech understanding tasks. The former is capable of detecting and recognizing keywords embedded in the utterance. The latter is to reject utterances that do not contain valid keywords and utterances that have low confidence scores. An important task in keyword spotting and utterance verification is the selection of an appropriate operating point or critical threshold to provide a desirable combination of *Type I error* (false rejection) and *Type II error* (false alarm).

Present *IODC* operators report that legitimate *IODC* users usually understand Mandarin, while prank callers often do not understand nor speak Mandarin. Hence to detect a prank caller, one may instead determine whether this caller understands and speaks Mandarin. This is realized in our prank call rejection system using both keyword spotting and utterance verification technologies. Upon receiving an incoming *IODC* call, the system issues an announcement in Mandarin asking the customer to say a keyword in Mandarin. If the system recognizes the keyword in the caller's response, it then switches the line to a human telephone operator. If not, the call is determined to be a prank call and the line is cut off automatically without being transferred to the operator.

In this paper, we report a recently developed prototype system for an application of prank call rejection using keyword spotting and utterance verification. The system issues an announcement in Mandarin which instructs *IODC* customers to pronounce a keyword in Mandarin. If the system recognizes the keyword in the response, it then switches the line to a

telephone operator. If not, the call is assumed to be a prank call and the line is cut off. This system was developed based on the fact that the Taiwanese or Chinese customers will understand the announcement but prank callers probably will not. In this paper, we mainly concern with a speech recognition technology used in the system.

The remainder of the paper is organized as follows. In Section 2 of this paper, we briefly describe system concepts. Phases of the development are discussed in Section 3. In Section 4, we describe a speech recognition technology used in the system. Experiment results are reported in Section 5. Finally, some conclusions are given in Section 6.

## 2. Concept of the System

Prank calls to Chunghwa Telecom IODC are made by natives of foreign countries who do not understand Mandarin. On the other hand, almost all customers of the service are Taiwanese. Hence, we designed the prank call rejection system as shown in Figure 1. After the keyword " **Chunghwa Telecom (中華電信)**" is announced in the system prompts, the IODC customers are connected with the operator only by repeating the keyword. The following shows an example of dialogue between a prank caller and the system.

User: (Call up system)

System: This is Chunghwa Telecom IODC service system. You are now connected to an automatic response system. Please say "**Chunghwa Telecom**" after the beep-tone, and we will connect you with the telephone operator (beep).

User: … (The system waits for few seconds.).

System: Please say "**Chunghwa Telecom**" once more after the beep-tone (beep).

User: #%&9?!

System: Sorry! Please call again.

### 3. Phases of the Development

The system was developed in the phases described as follows. A trial was made of incoming calls from the top 1 country where the prank call rates had been 80-90%.

**Phase 0:** Two telephone speech databases were setup to train and evaluate the proposed system. The first speech database (SDB1), used for training, consists of 400 phrases and short paragraphs that are chosen from TDB and read by 60 male and 40 female speakers. The second speech database (SDB2), used for testing, consists of 340 spontaneous utterances for IODC service uttered by 7 male speakers. And, there are 164 utterances containing valid keywords in SDB2.

**Phase 1:** A HMM-based keyword spotter and rejector were developed and integrated into the proposed system. A two-pass strategy was adopted consisting of recognition followed by verification. In the first pass, keyword spotting was performed to detect the position and its likelihood score of the possible keyword. In the second pass, for each keyword segmentation, a likelihood score was also obtained for the corresponding anti-keyword model. A confidence score based on a likelihood ratio test was then performed and the utterance was either accepted or rejected.

**Phase 2:** A selection of an appropriate operating point to provide a desirable combination of Type I error (false rejection) and Type II error (false alarm) were performed in this phase using SDB2. The experimental results demonstrate that 97.1% of prank calls were rejected while only 2.4% of customer calls were rejected.

**Phase 3:** The field-trial system was developed and has been in operation at the Chunghwa Telecom International Business Group since March 1998. A trial was only made of incoming calls from the top 1 prank call country. If the system recognizes the keyword, then it switches the line to a telephone operator. If not, the call is assumed to be a prank call and the line is cut off. Also, all calls were collected and will be used to improve the system performance.

## 4. Keyword Spotter and Rejector

In the Chunghwa Telecom *IODC* service system, the core technology module is the keyword spotter and rejector. Following keyword recognition, an input utterance was segmented and labeled as keyword and non-keyword hypotheses. Besides, their corresponding positions and HMM likelihood scores are also detected and calculated by the keyword spotter. Then, the rejector will verify whether the utterance is a prank call or not.

### 4.1. Keyword Spotter

In Chunghwa Telecom *IODC* service application, the expected utterance usually contains at most one keyword embedded in non-vocabulary speech. We inferred that performance could be significantly improved by imposing this single keyword constraint. To achieve this, we proposed a keyword-filler network. In this modified keyword spotter, only four kinds of utterances containing one valid keyword are allowed:

Type A: A single keyword
Type B: A single keyword followed by a non-keyword speech
Type C: A non-keyword speech followed by a single keyword
Type D: A single keyword embedded in non-keyword speech in both sides

In order to generate HMM models from *SDB1*, a segmental k-means training algorithm (Rabiner 1986) is used to optimize the likelihood of the observation sequence and the state sequence over all model parameters. To reduce the likelihood computation, subsyllabic units (Chen 1994), syllable initials and syllable finals, were used as basic HMM building blocks. Each initial and final model has 3 and 5 states, respectively. Overall there are 440 states for all the subsyllabic HMMs. A left-to-right HMM scheme with no skipped states was chosen for all the models.

### 4.2. Rejector

As a generalization to keyword spotting, utterance verification (*UV*) attempts to reject or accept an utterance based on a computed confidence score. This is particularly useful in situations where utterances are spoken without valid keywords or when significant confusion exists among keywords which may result in a high substitution error probability. *UV* is carried out by testing the *null hypothesis* that a specific keyword exists in a segment of speech $O$ versus the *alternative hypothesis* that the keyword is not present. Based on a likelihood ratio test, to accept or reject an utterance depends on whether the log likelihood ratio $LR(O \mid \Lambda)$ is higher than a specific verification threshold $\tau$ (here $\Lambda = \{ \lambda_t \}, \{ \lambda_a \}$). Sets of $\{ \lambda_t \}$ and $\{ \lambda_a \}$ are the models of the keyword and anti-keyword HMMs respectively.

Several different formulations for the alternative hypothesis have been proposed. Two formulations will be described in this section. The first choice is simply to use the general acoustic filler model $\lambda_f$ which is keyword independent. The likelihood for the alternative hypothesis is defined as $\log[p(O \mid \lambda_f)]$. The second choice for the alternative hypothesis is to introduce a keyword-specific anti-keyword model. There are many strategies for constructing such models, such as constructing additional keyword-specific anti-keyword models or using the likelihood of all competing models, $\{ \lambda_a \}$. The likelihood for the alternative hypothesis is defined as $\log[p(O \mid \lambda_a)]$. In this paper, we will only discuss the latter type since it does not need to train additional models and is easily constructed. The confidence measure is evaluated by the log likelihood ratio

$$LR(O \mid \Lambda) = \log[p(O \mid \lambda_t)] - \log[p(O \mid \lambda_a)], \qquad (1)$$

where $\log[p(O \mid \lambda_t)]$ is the likelihood for the null hypothesis.

An appropriate operating point is selected to provide a desirable combination of *Type I error* (false rejection) and *Type II error* (false alarm). Here, we chose an operating point to minimize the total error which is defined as the sum of false rejection and false alarm errors. An utterance is rejected if the test of the log likelihood ratio

$$LR(O \mid \Lambda) < \tau, \tag{2}$$

where $\tau$ is the operation point. This enables rejection of utterances which contain non-vocabulary words or noise. A general form of the likelihood of the alternative hypothesis based on anti-keyword model could be further formulated as

$$\log[\frac{1}{2}\exp\{\eta \log[p(O \mid \lambda_a)]\} + \frac{1}{2}\exp\{\eta \log[p(O \mid \lambda_f)]\}]^{\frac{1}{\eta}}, \tag{3}$$

where $\eta$ is a constant. Currently, experiments are conducted using the confidence measure function defined in equation (1) only.

## 5. Experiments

In order to evaluate the performance of our rejection scheme, experiments have been conducted with *SDB2* which consists of 340 spontaneous utterances for *IODC* service uttered by 7 male speakers. In *SDB2,* there are 164 utterances containing valid keywords and 176 utterances spoken without valid keywords. Figure 2 shows the two histograms for the keyword and non-keyword log likelihood ratio scores. Figure 3 shows overall system performance as a function of threshold. The FA (False Acceptance) is the rate of accepting prank calls and FR (False Rejection) is the rate of rejecting customer calls. We can control the rates of *Type I error* and *Type II error* with the threshold value. The operating point is designed to minimize the sum of false rejection and false alarm errors. The experimental results demonstrate that 97.1% of prank calls were rejected while only 2.4% of customer calls were rejected, as shown in Table 1. Figure 4 presents the ROC curve for the experiment. The underlying algorithm has very high probability of detection at very low false alarm rates, where the vocabulary size is only one. However, we also notice that as we increase the vocabulary size, the decrease in performance is evident in the experiments of the TL phone directory assistant task.

## 6. Conclusions

In order to reject prank calls for the Chunghwa Telecom *IODC* service, we developed the prank call rejection system using the technologies of keyword spotting and utterance verification. The system has been in operation at the Chunghwa Telecom International Business Group since March 1998. Over 90% of prank calls were successfully rejected in the first two weeks in the field-trial phase.

## Acknowledgments

The authors would like to thank Dr. J. T. Wang, Director of CHT-TL, Dr. J. H. Liang and Dr. B. S. Jeng, Deputy Director of CHT-TL, for their fruitful support. The authors also would like to thank Dr. K.-Y. Chang and Dr. C.-S. Liu for their invaluable advice and timely encouragement. We also thank the colleagues of the CHT-TL speech recognition group for their carrying out a certain part of the work described in this paper.

## References

Rohlicek, J. R., W. Russel, S. Roukos, H. Gish, "Continuous hidden Markov modeling for speaker-independent word spotting," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (Glasgow, Scotland), May 1989, pp. 627-630.

Rose, R. C., D. B. Paul, "A hidden Markov model based keyword recognition system," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (Albuquerque, New Mexico), April 1990, pp. 129-132.

Wilpon, J. G., L. R. Rabiner, C. H. Lee, E. R. Goldman, "Automatic recognition of keywords

in unconstrained speech using hidden Markov models," IEEE Trans. Acoust., Speech, Signal Processing, vol. 38, no. 11, pp. 1870-1878, November 1990.

Rahim, M. G., C. H. Lee and B. H. Juang, "Robust utterance verification for connected digits recognition," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing , 1995, pp. 285-288.

Kawahara, K., C. H. Lee and B. H. Juang, "Combining key-phrase detection and subword-based verification for flexible speech understanding," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1997, pp. 1159-1162..

Rabiner, L. R., J. G. Wilpon, and B. H. Juang, "A segmental k-means training procedure for connected word recognition based on whole word reference patterns," AT&T Tech. J., vol. 65, no. 3, pp. 21-31, May 1986.

Chen, J.-K., F. K. Soong, and L.-S. Lee, "Large vocabulary word recognition based on tree-trellis search," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (Adelaide, South Australia), April 1994, pp. II 137-140.

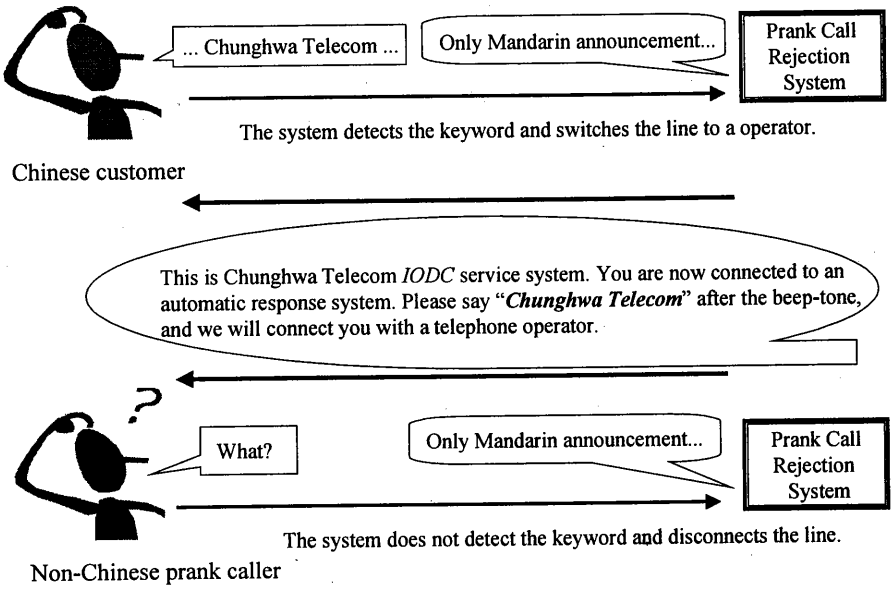


Figure 1. Dialogue between callers and a system.

| | 164 utterances with keywords | 176 utterances without keywords |
|---|---|---|
| Accept | Correct acceptance | *Type II error (5 utterances)* |
| Reject | *Type I error (4 utterances)* | Correct rejection |

Table 1.

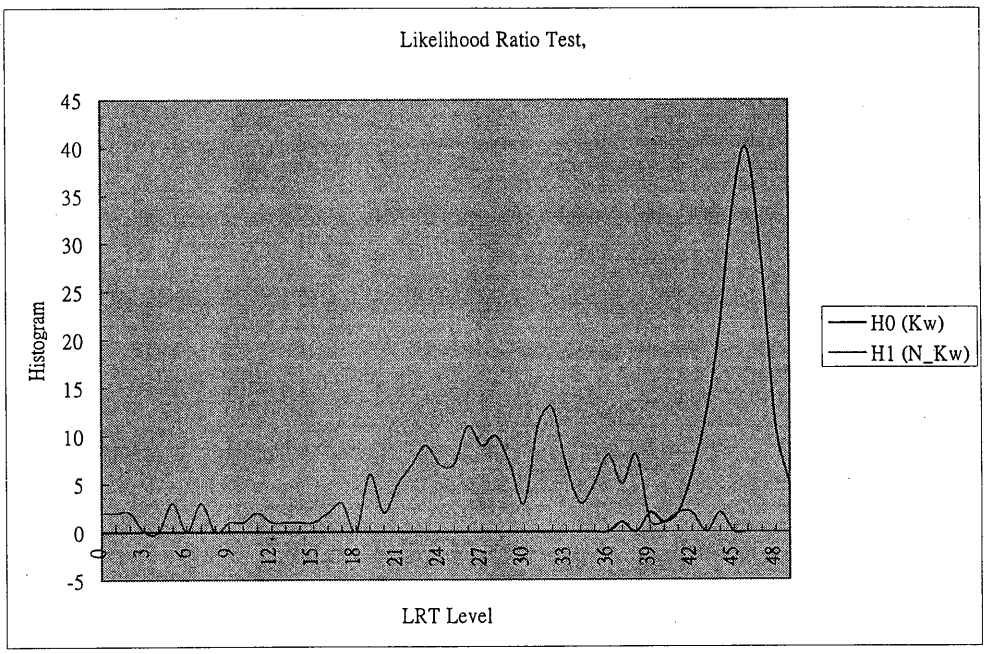


Figure 2. Histograms showing the distribution of the log likelihood ratio scores.

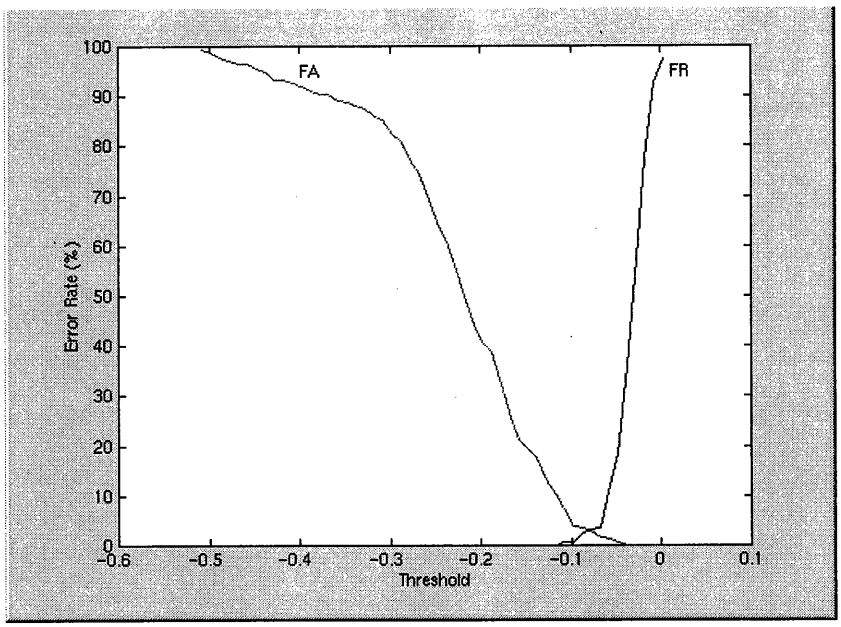


Figure 3. Error rate as a function of threshold.

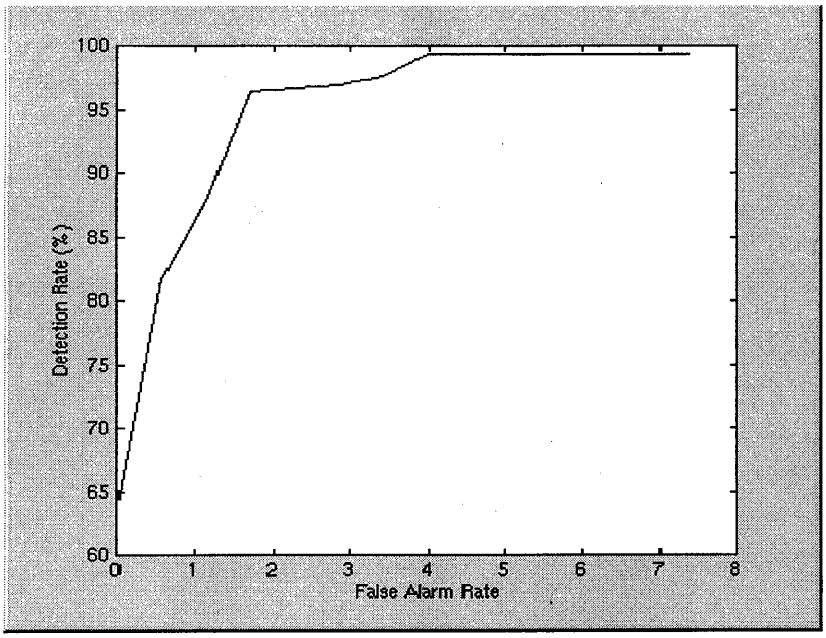


Figure 4. ROC curve.