Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2012, Article ID 548341, 8 pages
doi:10.1155/2012/548341

*Research Article*

# Optimization of Spoken Term Detection System

**Chuanxu Wang[1] and Pengyuan Zhang[2]**

[1] *Institute of Informatics, Qingdao University of Science and Technology, Qingdao 266061, China*
[2] *Thinkit Speech Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China*

Correspondence should be addressed to Chuanxu Wang, wangchuanxu_qd@163.com

Received 31 December 2011; Accepted 24 January 2012

Academic Editor: Baocang Ding



Generally speaking, spoken term detection system will degrade significantly because of mismatch between acoustic model and spontaneous speech. This paper presents an improved spoken term detection strategy, which integrated with a novel phoneme confusion matrix and an improved word-level minimum classification error (MCE) training method. The first technique is presented to improve spoken term detection rate while the second one is adopted to reject false accepts. On mandarin conversational telephone speech (CTS), the proposed methods reduce the equal error rate (EER) by 8.4% in relative.

## 1. Introduction

In recent years, there is an increasing trend towards the use of spoken term detection systems for real-world applications. In such systems, it is desirable to achieve the highest possible spoken term detection rate, while minimizing the number of false spoken term insertions. Unfortunately, most speech recognition systems fail to perform well when speakers have a regional accent. Particulary in China, the diversity of Mandarin accents is great and evolving.

Pronunciation variation has become an important topic. Normally, a confusion matrix is adopted to achieve higher recognition rate in speech recognition system. In [1], confusion matrix is adopted in spoken document retrieval system. Retrieval performance is improved by exploiting phoneme confusion probabilities. The work in [2] introduces an accent adaptation approach in which syllable confusion matrix is adopted. Similar approaches are discussed in [3].

The quality of confusion matrix has an obvious influence on the performance of spoken term detection. Based on traditional approaches, we propose an improved method to generate a phoneme confusion matrix.

MCE is one of the main approaches in discriminative training [4]. In [5], MCE is used to optimize the parameters of confidence function in large vocabulary speech recognition

system (LVCSR). The work in[6] introduces MCE into spoken term detection. In this paper, we present an improved MCE training method for calculating spoken term confidence.

The remainder of the paper is structured as follows: Section 2 introduces our baseline system. In Section 3, we discuss the phoneme confusion matrix based on confusion network. An improved MCE training method is presented in Section 4. In Section 5, the experiments are given and discussed, and finally Section 6 draws some conclusions from the proposed research.

## 2. Baseline System

In our baseline system, search space is generated based on all Chinese syllables, not specifically for spoken terms. Phoneme recognition is performed without any lexical constraints. Given a spoken input, our decoder outputs 1-best phoneme sequence. A phoneme confusion matrix is used to extract spoken terms.

The main steps of generating phoneme confusion matrix are listed as follows [2].

(1) Canonical pin-yin level transcriptions of the accent speech data should be obtained firstly.

(2) A standard Mandarin acoustic recognizer whose output is pin-yin stream is used to transcribe those accent speech data.

(3) With the help of dynamic programming (DP) technique, these pin-yin level transcriptions are aligned to the canonical pin-yin level transcriptions.

(4) Regardless of insertion and deletion errors, substitution errors are considered. Each pin-yin can be divided into two phonemes. Given a canonical phoneme $\mathrm{ph}_m$ and an aligning hypothesis $\mathrm{ph}_n$, we can compute confusion probability:

$$P(\mathrm{ph}_n \mid \mathrm{ph}_m) = \frac{\mathrm{count}(\mathrm{ph}_n \mid \mathrm{ph}_m)}{\sum_{i=1}^{N} \mathrm{count}(\mathrm{ph}_i \mid \mathrm{ph}_m)}, \tag{2.1}$$

where $\mathrm{count}(\mathrm{ph}_n \mid \mathrm{ph}_m)$ is the number of $\mathrm{ph}_n$ which is aligned to $\mathrm{ph}_m$. $N$ is the total phoneme number in dictionary.

With 1-best phoneme sequence and confusion matrix, similarities between phonemes are computed. For each spoken term, corresponding phonemes will be searched from pronunciation dictionary firstly. Then, sliding window is used to align phonemes of spoken term and 1-best phoneme sequence. The step of sliding window is set to two because there are two phonemes in each syllable in Chinese. An example of searching "gu zhe" is given in Figure 1.

Given a term $\varphi_1$, $\varphi_2$ is the aligning 1-best phoneme sequence. Then, similarity between them is denoted as $\mathrm{Sim}(\varphi_1, \varphi_2)$:

$$\mathrm{Sim}(\varphi_1, \varphi_2) = \frac{1}{N} \log\left( \prod_{i=1}^{N} P(\beta_i \mid \alpha_i) \right), \tag{2.2}$$

where $\alpha_i$ and $\beta_i$ are the $i_{\mathrm{th}}$ phoneme of $\varphi_1$ and $\varphi_2$, respectively, $N$ is the number of phonemes of $\varphi_1$.

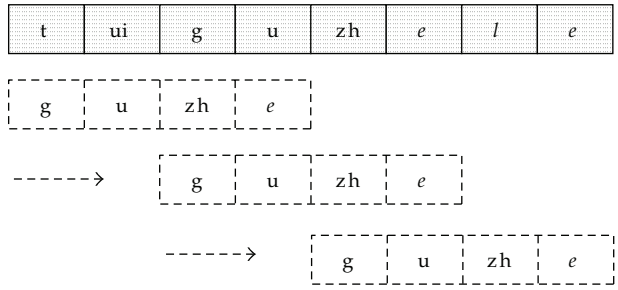


**Figure 1:** Extraction of "gu zhe."

Spoken term rate gets a significant improvement with the help of confusion matrix. But at the same time, false accepts have been increased too. Effective confidence measure should be adopted to reject false hypotheses. In this paper, word confidence is calculated with catch-all model [5]. A confidence score for a hypothesized phoneme $\mathrm{ph}_i$ is estimated by

$$\mathrm{CM}(\mathrm{ph}_i) = \frac{1}{e[i]-b[i]+1}\sum_{n=b[i]}^{e[i]} \log p\left(q^{(n)}\middle| o^{(n)}\right) = \frac{1}{e[i]-b[i]+1}\sum_{n=b[i]}^{e[i]} \log \frac{P\left(o^{(n)}\middle| q^{(n)}\right)P\left(q^{(n)}\right)}{P\left(o^{(n)}\right)}, \tag{2.3}$$

where $b[i]$ is the start time of $\mathrm{ph}_i$ and $e[i]$ is the end time. $q^{(n)}$ represents Viterbi state sequence.

Deriving word level scores from phoneme scores is a natural extension of the recognition process. We adopted the arithmetic mean in logarithmic scale. Spoken term confidence $\mathrm{CM}_{\mathrm{pos}}$ is defined as

$$\mathrm{CM}_{\mathrm{pos}}(w) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{CM}(\mathrm{ph}_i), \tag{2.4}$$

where $m$ is the number of phonemes in $w$.

## 3. Confusion Matrix Based on Confusion Network

Just as the above description, confusion matrix is generated from 1-best hypothesis. However, there is a conceptual mismatch between decoding criterion and confusion probability evaluation. Given an input utterance, a Viterbi decoder is used to find the best sentence. But it does not ensure that each phoneme is the optimal one. In this paper, we propose an improved method of generating confusion matrix. Instead of 1-best phoneme hypothesis, we get hypotheses from confusion network (CN) [7].

Just as Figure 2 describes, CN is composed of several branches. For schematic description, we give top 4 hypotheses in each branch. Corresponding canonical pin-yin stream is also presented in Figure 2.

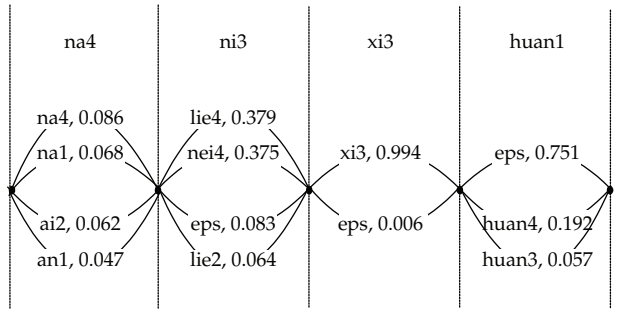


**Figure 2:** An example of confusion network.

Experimental show that syllable error rate (SER) of CN is far lower than that of 1-best sequence. Base on this point, we believe that CN provides us more useful information. In this paper, we attempt to use n-best hypotheses of each branch. Firstly, canonical pin-yin level transcriptions are formatted into a simple CN. Then, recognizer output voting error reduction (ROVER) technology is adopted to align two CNs. At last, we select special branches to generate confusion matrix. Given a canonical phoneme $\mathrm{ph}_m$, only branches including $\mathrm{ph}_m$ are considered. A sequence of class labels $\alpha(k)$ is defined as

$$\alpha(k) = \begin{cases} 1 & \text{if } \mathrm{ph}_m \in \text{the } k\text{th Branch}, \\ 0 & \text{if } \mathrm{ph}_m \notin \text{the } k\text{th Branch}. \end{cases} \tag{3.1}$$

Then, (2.1) can be rewritten as

$$P(\mathrm{ph}_n \mid \mathrm{ph}_m) = \frac{\sum_{k=1}^{C} \alpha(k)\mathrm{count}(\mathrm{ph}_n \mid \mathrm{ph}_m)}{\sum_{i=1}^{N}\sum_{k=1}^{C} \alpha(k)\mathrm{count}(\mathrm{ph}_i \mid \mathrm{ph}_m)}, \tag{3.2}$$

where $C$ is the number of branches in CNs of training data, $N$ is the number of phonemes in dictionary.

Another optional method is also attempted in this paper. Max probability rule can be applied in calculating confusion probability. The branches with maximum probability $\mathrm{ph}_m$ are considered. We define

$$\beta(k) = \begin{cases} 1 & \text{if } \mathrm{ph}_m \text{ is a phoneme with maximum probability in the } k\text{th Branch}, \\ 0 & \text{others.} \end{cases} \tag{3.3}$$

Then, (3.2) can be rewritten as

$$(\mathrm{ph}_n \mid \mathrm{ph}_m) = \frac{\sum_{k=1}^{C} \beta(k)\mathrm{count}(\mathrm{ph}_n \mid \mathrm{ph}_m)}{\sum_{i=1}^{N}\sum_{k=1}^{C} \beta(k)\mathrm{count}(\mathrm{ph}_i \mid \mathrm{ph}_m)}. \tag{3.4}$$

## 4. MCE with Block Training

The work in [5] proposed a word-level MCE training technique in optimizing the parameters of the confidence function. In [6], a revised scheme is implemented under spoken term scenario. In this paper, we attempt to improve the MCE training methods proposed in [6].

According to the update equations in [6], sequential training is used to update parameters. That is to say, the parameters of triphones are modified with each training sample. It is not matched well with optimization method of MCE. We adopt block training method instead. The parameters are modified with all averaged samples at once. The weighted mean confidence measure of $W$ is defined as

$$\mathrm{CM}(W) = \frac{1}{N_w}\sum_{i=1}^{N_w}\left(a_{\mathrm{ph}_i}\mathrm{CM}\left(\mathrm{ph}_i\right) + b_{\mathrm{ph}_i}\right). \tag{4.1}$$

Procedures of block training are listed as follows.

(1) Misclassification measure is defined as

$$d(W) = \left(\mathrm{CM}(W) - C\right) \times \mathrm{Sign}(W), \tag{4.2}$$

where $C$ is confidence threshold, $\mathrm{Sign}(W)$ is defined as

$$\mathrm{Sign}(W) = \begin{cases} 1 & \text{if } W \text{ is incorrect,} \\ -1 & \text{if } W \text{ is correct.} \end{cases} \tag{4.3}$$

(2) A smooth zero-one loss function is given by

$$l(W) = \frac{1}{1 + \exp\left(-\gamma d(W)\right)}. \tag{4.4}$$

(3) The parameter estimation is based on the minimization of the expected loss which, for a training sample of size $M$, is defined as

$$l\left(\overline{W}\right) = E(l(W)) = \frac{1}{M}\sum_{j=1}^{M} l(W_j). \tag{4.5}$$

Generalized probabilistic descent (GPD) algorithm is used to minimize the loss function $l(\overline{W})$ [8]:

$$\frac{\partial l\left(\overline{W}\right)}{\partial a_{\mathrm{ph}_i}} = \frac{1}{M}\sum_{j=1}^{M}\frac{\partial l(W_j)}{\partial a_{\mathrm{ph}_i}} = \frac{\gamma}{M_{\mathrm{ph}_i}}\sum_{j=1}^{M_{\mathrm{ph}_i}}\frac{1}{N_j}K(W_j)\mathrm{CM}_j(\mathrm{ph}_i),$$

$$\frac{\partial l\left(\overline{W}\right)}{\partial b_{\mathrm{ph}_i}} = \frac{1}{M}\sum_{j=1}^{M}\frac{\partial l(W_j)}{\partial b_{\mathrm{ph}_i}} = \frac{\gamma}{M_{\mathrm{ph}_i}}\sum_{j=1}^{M_{\mathrm{ph}_i}}\frac{K(W_j)}{N_j}, \tag{4.6}$$

$$\frac{\partial l\left(\overline{W}\right)}{\partial C} = \frac{1}{M}\sum_{j=1}^{M}\frac{\partial l(W_j)}{\partial C} = \frac{-\gamma}{M}\sum_{j=1}^{M}K(W_j),$$

where $M_{\mathrm{ph}_i}$ is the number of samples that contain the phoneme $\mathrm{ph}_i$, $N_j$ is the number of phonemes of $W_j$, $\mathrm{CM}_j(\mathrm{ph}_i)$ is the confidence of $\mathrm{ph}_i$ in $W_j$. However $k(W_j)$ is defined as

$$k(W_j) = l(W_j)(1 - l(W_j))\mathrm{Sign}(W_j). \tag{4.7}$$

At last, we get the revised update equations as

$$\widetilde{a}_{\mathrm{ph}_i}(n+1) = \widetilde{a}_{\mathrm{ph}_i}(n) - \varepsilon_n\frac{\partial l\left(\overline{W}\right)}{\partial a_{\mathrm{ph}_i}}\exp\left(\widetilde{a}_{\mathrm{ph}_i}(n)\right),$$

$$b_{\mathrm{ph}_i}(n+1) = b_{\mathrm{ph}_i}(n) - \varepsilon_n\frac{\partial l\left(\overline{W}\right)}{\partial b_{\mathrm{ph}_i}}, \tag{4.8}$$

$$C(n+1) = C(n) - \varepsilon_n\frac{\partial l\left(\overline{W}\right)}{\partial C}.$$

## 5. Experiments

We conducted experiments using our real-time spoken term system. Acoustic model is trained using train04, which is collected by Hong Kong University of Science and Technology (HKUST).

### 5.1. Experimental Data Description

The test data is a subset of development data (dev04), which is also collected by HKUST. Total 20 conversations are used for our evaluation. 100 words are selected as the spoken term list, including 75 two-syllable words and 25 three-syllable words.

Confusion matrixes adopted in this paper are generated using 100-hour mandarin CTS corpus. The word-level MCE training set is a subset of train04 corpus. 865667 terms are extracted for the training, including 675998 false accepts and 189669 correct hits.

**Table 1:** SER of CN and 1-best pin-yin sequence.

| | SER | Pruning beam |
|---|---|---|
| 1-best pin-yin | 75.0% | — |
| CN | 71.0% | 0.5 |
| CN | 51.9% | 0.2 |
| CN | 35.3% | 0 |

**Table 2:** Recognition rates of different phoneme confusion matrixes.

| Confusion matrix | Recognition rates |
|---|---|
| 1-best recognition result | 77.3% |
| CN | 80.9% |
| CN + maximum probability | 82.0% |

### *5.2. Experiment Results*

The detection error tradeoff (DET) is used in this paper to evaluate the performance of spoken term. The false acceptation (FA) rate fits the case in which an incorrect word is accepted, and the false reject (FR) fits the case of rejecting the correct word:

$$\begin{aligned} \text{FA} &= \frac{\text{num. of incorrect words labelled as accepted}}{\text{num. of incorrect words}}, \\ \text{FR} &= \frac{\text{num. of correct words labelled as rejected}}{\text{num. of keywords} * \text{hours of testset} * C}, \end{aligned} \tag{5.1}$$

where $C$ is a factor which scales the dynamic range of FA and FR on the same level. In this paper, $C$ is set to 10. Recognition rates (RA) are also computed. It can be obtained as:

$$\text{RA} = \frac{\text{num. of correct words labelled as accepted}}{\text{total num. of recognized words}}. \tag{5.2}$$

In order to assess how CN gives more information than 1-best pin-yin sequence, the syllable error rates (SERs) of both CN and pin-yin sequence are given in Table 1. SER of CN drops significantly with the reduction of pruning beam.

Table 2 summarizes recognition rates of different confusion matrixes. With the $n$-best hypotheses of CN, recognition rates are improved obviously. Then maximum probability rule is applied, and the recognition rate arrives 82.0%.

To evaluate the performance of methods proposed in this paper, EERs of different methods are listed in Table 3.

As we can see from Table 3, the improved confusion matrixes provide obviously EER reduction of up to 3.9% in relative. MCE with block training is superior to sequential training, relative 1.7% EER reduction is achieved. When two methods are used at the same time, we get a further improvement, 8.4% relative reduction compared with the baseline system.

**Table 3:** EER of different methods.

| Methods | EER |
|---|---|
| Baseline | 48.8% |
| CN | 47.5% |
| CN + maximum probability | 46.9% |
| MCE with sequential training | 46.2% |
| MCE with block training | 45.4% |
| CN + maximum probability + MCE with block training | 44.7% |

## 6. Conclusions

In order to describe how the accent-specific pronunciation differs from those assumed by the standard Mandarin recognition system, the phoneme confusion matrix is adopted. Different from traditional algorithm, confusion network is applied in generating confusion matrix. It improves the recognition rate of spoken term system. Moreover, a revised MCE training method is presented in this paper. Experiments prove that it performs obviously better than the sequential training.

## References

[1] N. Moreau, H.-G. Kim, and T. Sikora, "Phonetic confusion based document expansion for spoken document retrieval," in *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP '04)*, Jeju Island, Korea, October 2004.

[2] M. Liu, B. Xu, T. Huang, Y. Deng, and C. Li, "Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling," in *Proceedings of the the IEEE Interntional Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, pp. 1025–1028, June 2000.

[3] J. Gao, Q. Zhao, Y. Yan, and J. Shao, "Efficient system combination for syllable-confusion-network-based Chinese spoken term detection," in *Proceedings of the 6th International Symposium on Chinese Spoken Language Processing (ISCSLP '08)*, pp. 366–369, December 2008.

[4] B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, no. 12, pp. 3043–3054, 1992.

[5] S. Abdou and M. S. Scordilis, "Beam search pruning in speech recognition using a posterior probability-based confidence measure," *Speech Communication*, vol. 42, no. 3-4, pp. 409–428, 2004.

[6] J. Liang, M. Meng, X. Wang, P. Ding, and B. Xu, "An improved mandarin keyword spotting system using MCE training and context-enhanced verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, pp. I1145–I1148, May 2006.

[7] T. Hori, I. L. Hetherington, T. J. Hazen, and J. R. Glass, "Open-vocabulary spoken utterance retrieval using confusion networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 4, pp. IV73–IV76, April 2007.

[8] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Cambridge University, Cambridge, UK, 2004.