REJECTION BASED ON A *POSTERIORI* PROBABILITY ESTIMATED BY MLP WITH APPLICATION FOR MANDARIN VOICE DIALER ON ASIC

Lin Zhong, Jia Liu and Runsheng Liu

Dept. of Electronic Engineering
Tsinghua University, Beijing 100084
P.R.China
zhongl@hannah.ee.tsinghua.edu.cn

Abstract

High performance mandarin voice dialer is much more difficult than its English counterpart to achieve, especially on inexpensive hardware as ASIC. One way to improve its performance is to incorporate rejecters into the system. In our study, an MLP based postprocessor, an *a posteriori* probability estimator, is applied after HMM Viterbi recognizing. Poor utterances, which are recognized by HMMs but have low *a posteriori* probability, are then rejected. Rejecting 4.9% of all the testing utterances, the MLP rejecter boost the HMM-based system's single digit accuracy from 97.1% to 99.6% for the Mandarin voice dialer, a ten-syllable speaker independent task. The performance is better than those of rejection based on linear discrimination, anti-digit models or likelihood ratio.

1. INTRODCUTION

Rejection is widely adopted mainly for two reasons, 1) to exclude utterances that are out of vocabulary(OOV), and 2) to ignore utterances that are poorly recognized. In the voice dialer task, both issues are significant. To be user friendly, the voice dialer have to be robust as well as accurate. However, in this paper, we only focus on the second issue, i.e., using rejection to improve the system accuracy, although the MLP rejecter is also readily applicable for rejecting OOV utterances, as discussed in section 7.

Various methods have been developed to achieve rejection, most in the context of keyword spotting. Wilpon et al[1] used garbage or filler models to model OOV utterance. Sukkar et al[2][3] adopted linear discrimination for rejection. Rahim et al[4] developed the concept of anti-keyword to verify whether an utterance contains a putative keyword, in their application, English digits. In [5], Villarrubia and Acero propose the use of an affine transformation which is essentially linear. Especially, Mathan and Miclet[6] presented an MLP taking in the trace of HMMs for rejection extraneous input. The MLP is trained to perform binary classification, which takes the advantage of simplicity. While in our study, the MLP is extended to take in the trace of all the current models and trained to estimate the *a posteriori* probability[7].

In section 2.,details about the mandarin voice dialer and how we evaluate the rejecter are presented. Then section 3 and section 4 describes the HMM-based recognizer and details the MLP rejecter, respectively. Rejecters based on other paradigms are described in section 5 and are compared with the MLP rejecter in section 6. And our conclusions are offered in section 7.

2. MANDARIN VOICE DIALER

With voice dialers, users can dial by speaking out the phone number or the name pre-stored. While the later only requires speaker dependent and isolated speech recognition, the former is rather a technical challenge[8]. Unlike digits in other language such as English, Mandarin digits are in fact syllables, and high confusion here exists. In this respect, the task resembles the English alphabet recognition[9], while on a smaller scale. Moreover, due to the storage and computation limits imposed by the hardware when applied to consumer electronic products, voice dialers can not adopt most of the methods developed to improve English alphabet recognition.

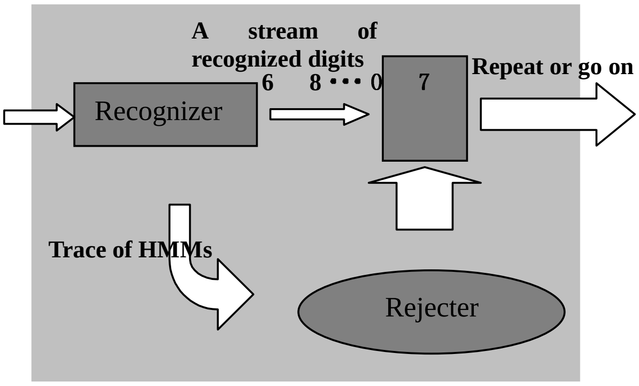

Figure 1. The Mandarin Voice Dialer System

The mandarin voice dialer currently under development in the Speech Technology and ASIC Realization Group in Tsinghua University, is a system based on HMMs and is expected to fulfill speaker independent continuous mandarin digits recognition on ASIC. Despite the technical difficulty, the application requires very high accuracy and robustness because false acceptance is very costly for dialing. On one hand, researches have been conducted to find methods that boost the recognition accuracy while demanding relatively little resource. On the other hand, it is very desirable to incorporate a rejecter into the dialer so that any single digit from a string can be rejected if it is recognized with little confidence, and the user can then repeat that digit instead of the whole string. Working together with the rejecter, recognizers that is not so accurate without rejection become viable and friendly. The system is showed in figure 1.

According to the task involved, we evaluate rejecters with the rejecting rate and accuracy after rejection as defined below.

Rejecting Rate
RR = number of rejected utterances / number of testing utterances.
Accuracy after Rejection
AR = number of correctly recognized utterances / number of accepted utterances.

Rejecters' performance can be compared by the curve of *AR* vs. *RR*.. The higher the rejection rate, the more the user has to repeat, while the higher the accuracy, the less errors the user is expected to encounter. Usually, with other conditions fixed, the higher the *RR*, the higher the *AR*. A tradeoff between the two devils has to be made carefully.

3. HMM-BASED RECOGNIZER

The recognizer is based on HMMs. 10LPCCs , energy and their one order differentials, i.e., 22 features per frame, are used as speech features. LPCCs instead of MFCCs are adopted because they demand much less computation resource and thus facilitate fulfillment on ASIC.

As the mandarin digits are syllables, we choose to model them by whole word models, i.e., whole syllable models. The HMM for each digit has 6 state and each state has 7 Gaussian densities. The HMM-based recognizer achieve single digit accuracy without rejection of 97.1%. Such performance is remarkable in view of the high confusion among such syllables. However, it isn't satisfactory if the system is used for voice dialer, in which usually a series of digits have to be recognized.

4. MLP-BASED REJECTER

Multi-layer Perceptrons have long been valued for their learning ability. When certain conditions are warranted, they perform as the *a posteriori* probability estimator[7]. In our application, the MLP is trained by the trace generated by all the HMMs, in our case, 10 HMMs for the Mandarin digits. After HMMs are trained, all the training data are Viterbi aligned with every HMM to generate the trace. Then, the trace generated by all the HMMs for an utterance are used to train the MLP, which, for training, has been assigned 1 as the ideal output for the corresponding speech class and 0 for the others.

4.1 Rejecting Criterion

If there are enough parameters in the MLP and the training process isn't trapped by a local minimum, the MLP outputs the *a posteriori* probability of the speech class conditioned on the input trace[7]. The conditions assumed, the MLP's output node corresponding speech class *k* presents $O_k = P(\omega_k \mid \overrightarrow{Tr})$, where ω_k denotes the speech class of k, and $\overrightarrow{Tr}$ denotes the trace generated by all the HMMs for an utterance.

While most other rejection criteria are based on likelihood ratios, ours is based on the *a posteriori* probability estimated by the MLP. After the HMM-based recognizer report the identity of a digit in a string, for example, digit *k*, the MLP takes in the trace generated by the HMMs and output the *a posteriori* probability $P(\omega_k \mid \overrightarrow{Tr})$. Then the system decides whether to reject the utterance by comparing $P(\omega_k \mid \overrightarrow{Tr})$ with a prefixed value ε ,i.e., reject the utterance if $P(\omega_k \mid \overrightarrow{Tr}) < \varepsilon$.

Increasing ε will increase *RR*.. Indeed, varying the ε ,the rejecter will operate at different point on the *RR-AR* curve. Therefore, experiments have to be done to find the ε that ensures the best operating point for the task.

4.2 Features for Rejection

In[6], the trace of HMMs, i.e. duration and average frame in the state is proved to be informative. Therefore, we adopted them as features for rejection. The trace consists of two parts, 1) relative duration in each state of the HMM, and 2) average frame features assigned to each state of the HMM. As there are 22 features per frame and 6 state per HMM, the MLP has to take in 138 ∗ *N* features, where *N* denote the number of current models. If all the other models are taken as current models, i.e., the MLP has to take the traces generated by all the HMMs, the storage and computation demand is formidable for simple hardware fulfillment. Therefore, the input features should be reduced. Figure 2. shows the *RR-AR* curves when different components of the trace are combined.

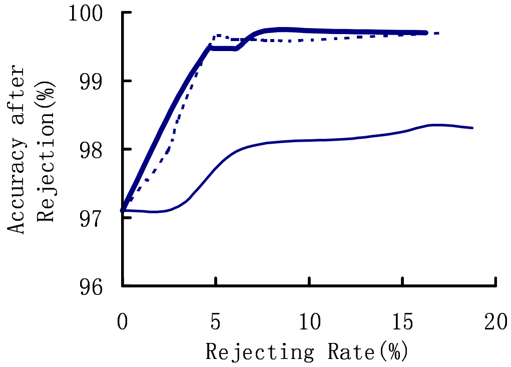

Figure 2. *RR-AR* Curves for MLP rejecters with 24 hidden Neurons

The thick curve in Figure 2 is derived with all the trace as the input to the MLP rejecter, while the solid thin curve with only the relative duration as input, and the dashed curve with the relative duration and the average cepstrum coefficient and energy frame features assigned to the state, i.e., ignoring their differentials. Apparently, the rejecter with the full trace achieve the best performance. However, the rejecter which ignores the differentials achieve nearly the same performance but requires nearly half of the storage and computation resource. Then, the later is adopted in our system and compared with other methods in this paper.

5. REJECTERS BASED ON OTHER PARADIGMS

5.1 Linear Discrimination

The rejecter based on linear discrimination is similar in nature to that in [2] and [5]. In this study, the trace generated by all the current HMMs is linearly transformed and then normalized by the sigmoid function. That is,

$$y_k = \overrightarrow{W_k}\,\overrightarrow{Tr} + b_k \text{ and } O_k = \frac{1}{1+e^{-y_k}}$$

where *k* denote the speech class, $\overrightarrow{W_k}$ and b_k are the corresponding weight vector and bias, respectively. The linear discrimination rejecter is trained and works in a very similar way to that of the MLP rejecter. Indeed, the linear discriminants are just single-layer perceptrons. However, the MLP outperforms the linear discrimination in the ability to approximate arbitrary distributions. Anyway, later it will be shown that linear discrimination rejecters defeat rejecters based on likelihood ratio and anti-word models easily , and performs only a little worse than MLP rejecters.

5.2 Likelihood Ratio

The HMM recognizer readily provides the rejecter the likelihood scores of an utterance for the current speech classes. Usually the likelihood ratio is adopted for rejection. The simplest version is presented here as the benchmark rejecter. An utterance, which is recognized as from speech class *k,* is rejected if

$$\frac{P(\overrightarrow{X}\,|\,\omega_k)}{\max_{j\neq k} P(\overrightarrow{X}\,|\,\omega_j)} < \varepsilon,$$ where $\overrightarrow{X}$ denotes the utterance feature vector

5.3 Anti-word Models

Anti-digit models[4] are trained for each mandarin digit in our study, using utterances that belong to the competing digits. The anti-digit model has the same number of states and each state has the same number of Gaussian densities as those of the digit models. An utterance, which is recognized as from speech class *k*, is rejected if

$$\frac{P(\overrightarrow{X}\,|\,\omega_k)}{P(\overrightarrow{X}\,|\,\overline{\omega_k})} < \varepsilon,$$ where $\overline{\omega_k}$ denotes the anti-digit model for speech class *k.*

6. EXPERIMENT RESULTS

We trained and tested the rejecters with a speech database consisting of utterances by 160 speakers, 80 males and 80 females. The testing method of Jack-knife is adopted to make full use of the database.

Figure 3. shows the *RR-AR* curves for rejecters based on a MLP with 12 hidden neurons(the thick line), likelihood ratio(the dashed thin line) and anti-digit models(the solid thin line).Obviously, the MLP-based rejecter easily outperform the other two.

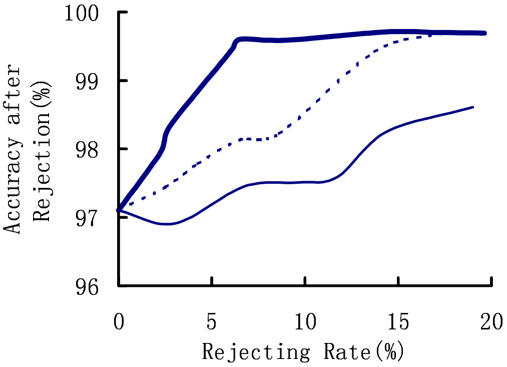

Figure 3. *RR-AR* curves for different rejecters

At first, the bad performance by anti-digit models based rejecter seems rather perplexing. Then we realized that the asymmetry in confusion may count for it. There are two kinds of asymmetry, 1) some digit is severely confused with others while other digits is seldom mistaken, i.e., recognition errors are mainly ascribed to several confusion pairs; 2) in certain confusion pair, always, one is mistaken as the other while the later seldom is mistaken as the former. Because of such asymmetries, the anti-digit model, which is trained by all the competing digits, fails to provide the likelihood that the utterance comes from other speech classes as accurate as the likelihood that the utterance come from the current speech class , which is provided by the much more delicate digit model. We're looking forward to training the anti-digit model using utterances from carefully chosen competing classes. Good performance may be obtained then.

Figure 4. shows the *RR-AR* curves for rejecters based on MLP with 24 hidden neurons(solid thin line),MLP with 12 hidden neurons(thick line), and linear discrimination(dashed thin line).

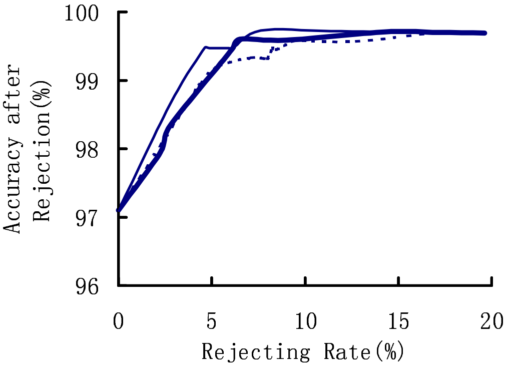

Figure 4. *RR-AR* curves for MLP rejecters and the linear discrimination rejecter

Clearly, the rejecter based on MLP with 24 hidden neurons performs better than the other two, especially the linear discrimination rejecter. By rejecting 4.9% of the testing utterances, it boost the recognition accuracy from 97.1% to 99.6%. However, although the rejecter based on MLP with 12 hidden neurons achieves slightly worse performance, it only requires nearly half the storage and computation resource as the one based on MLP with 24 hidden neurons does. In this regard, if further tradeoff between resource and performance should be made, the rejecter based on the MLP with 12 hidden neurons is recommended. It's worth mentioning that the linear discrimination rejecter achieve rather remarkable performance, much better than those of anti-digit model based and likelihood ratio based rejecters. Further discussion is available in the next section.

7. DISCUSSION AND PROSPECTS

7.1 Tradeoff between Resource and Performance

Tradeoff between resources and performance is vital for the hardware realization of speech recognition, especially for fulfillment on ASIC for consumer electronic products. In addition to the performance depicted in Figure 3 and 4, the following table details the resource demand by the mentioned rejecters in terms of the number of the parameters in the rejecter.

Table 1. Number of parameters in rejecters

Rejecters	Num. Of Parameters
MLP with 12 Hidden Neurons(small)	8,782
MLP with 24 Hidden Neurons(big)	17,554
Linear Discrimination	7,210
Anti-digit Models	18,900
Likelihood Ratio	0

Judging by table 1 and Figure 3 and 4, when resource is very limited, the simple likelihood ratio rejecter is a decent solution. And when resource is ample, the rejecter based on the big MLP is preferable for it has the best rejecting performance. The rejecter based on the small MLP offers the best tradeoff between resource and performance, and is most suitable for our mandarin voice dialer on ASIC. The MLP based rejecters defeat the linear discrimination rejecter only slightly in performance, but the former provide us with flexibility and more choices. Just changing the network scale by varying the number of hidden neurons, we can get different tradeoffs between resource ad performance. Moreover, the probabilistic interpretation for the output of the MLP rejecter also facilitates more delicate statistical rejection in further studies. We are now conducting researches into further exploiting the statistical characteristics of the MLP rejecters. One of the many possibilities is using the MLP for OOV rejection.

7.2 Using MLP rejecters for OOV Rejection

The extension from rejecting poorly recognized utterances to rejecting OOV utterances is direct using the MLP rejecter, as is currently underway in our group. Filler or garbage hidden Markov models are trained to model the OOV utterances. Several more nodes are added to the MLP to represent OOV utterances that belong to different categories. The MLP rejecter is trained by the trace of digit utterances and OOV utterances generated by digit models and filler models. Then a recognized utterance is rejected as OOV if the output for any of the OOV categories exceeds a thresh and is rejected as mis-recognized if the output for the reported digit falls below another thresh.. Preliminary researches have proved that the MLP rejecter is promising. Moreover, we are trying to incorporate other rejection paradigm into the *a posteriori* probability estimation frame.

8. Acknowledgement

This project is supported by National Natural Science Funds (Item:69772020), National 863 High Technology Projects (Contract No. 863-512-9805-10).

9. Reference

[1]Wilpon, J.G., Rabiner, L.R., Lee, C.-H. & Goldman, E.R.(1990), Automatic Recognition of Keywords in Unconstrained Speech Using HMM's, *IEEE Trans. On ASSP*, vol.38,n.11,1870-1878.

[2]Sukkar, R.A. & Wilpon, J.G.(1993), A two pass classifier for utterance rejection in keyword spotting, *IEEE Proc. ICASSP*,vol.2,451-454.

[3]Sukkar, R.A.(1994), Rejection for connected digit recognition based on GPD segmental discrimination, *IEEE Proc. ICASSP*,vol.1,393-396.

[4]Rahim, M.G., Lee, C.-H. & Juang, B.-H.(1997), Discriminative Utterance Verification for Connected Digits Recognition, *IEEE Trans. On SAP*, vol.5, n.3, 266-277.

[5]Villarrubia, L. and Acero, A.(1993), Rejection techniques for Digit Recognition in Telecommunication Applications, *IEEE Proc. ICASSP*,vol.2,455-458.

[6]Mathan, L. & Miclet, L.(1991), Rejection of extraneous input in speech recognition application using MLP's and the trace of HMM's, *IEEE Proc. ICASSP*, vol.1, 93-96.

[7]Richard, M.D. and Lippmann，R.P. (1991), Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities, *Neural Computation*,vol.3, pp.461-483.

[8]Gu, L. and Liu, Runsheng(1997), Mandarin Digit Speech Recognition: State of the Art, Difficult Points Analysis and Methods Comparison, *Journal of Circuits and Systems*, In Chinese, vol.2, no.4, pp.32-39.

[9]Loizou, P.C. and Spanias, A.S.(1996), High_performance Alphabet Recognition, *IEEE Trans. SAP*, vol.4,no.6 , pp.430-445.