ABSTRACT

Since telephone is the only ubiquitous communications device in current world, it is the largest potential application field for speech techniques. Telephony speech recognition is a core technique for such telephone-based speech applications. It is well known that the bandwidth of telephone line is limited to 300~3400Hz and there are many inherent variations within the telephone network. All these make speech recognition over telephone a more difficult task compared to its desktop pairs. Additionally, due to the freely speaking style required by real applications and the diverse background environment, a perfect laboratory system may become very vulnerable in real world. So the robustness is the life-and-death issue for such commercial systems. In this paper, we will introduce our recent progresses on improving the performance for a Mandarin telephony speech recognition system. Our improvements include a more robust and straightforward feature extraction block for telephony speech and a novel dynamic channel compensation algorithm. And then we will focus our discussion on the strategy of dealing with out-of-vocabulary (OOV) utterances. Through all these amendments, the system’s performance obviously improves in real applications.

1. INTRADUCTION

The framework of our telephony speech recognition system is depicted in figure (1).

Fig 1. System framework on Acoustic level

The state of the art speech recognition systems are capable of highly accurate performance for particular tasks within well-defined operational environments. For these ideal environments, the hatched blocks in figure 1 may be not essential ones. But for real application-oriented systems, such as telephone-based systems, the channel distortion and complicated background could not be neglected. Some compensation techniques could be used to enhance the noise contaminated and channel distorted speech. Alternatively the acoustic model could be modified to reflect the prevailing conditions using model adaptation techniques. To provide robustness for unexpected sound events, which is inevitable in every real application, adjustments should be made to reject the out-of-vocabulary words. In the following sections of this paper, we will present our improvement under this framework.

2. ROBUST FEATURE EXTRACTION

Speech signal is the vocal tract’s response corresponding to the excitation from the lungs. So the vocal tract parameter set is the most suitable feature for the speech recognizer. Until now, the cepstrum parameter set, which describes the envelope of vocal tract frequency response, gives the best discrimination between different speech events. Many researches have been done to introduce the psychophysical processes of human hearing into the cepstrum analysis, such as Mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive analysis (PLP).

It is well known that Chinese is a tonal language and tone, the trajectory of pitch, plays a critical role in Chinese. So pitch is an important feature for Mandarin recognition. However, due to the bandwidth limitation and channel distortion exists in the communications network, pitch extracting over telephone lines becomes a very difficult work. Existing pitch extraction algorithm could hardly fulfill the precision requirement for the recognizer and the real-time requirement for the telephone applications at the same time. We improve short time analysis for the fundamental frequency and integrate it with the common MFCC extraction procedure. Additional computation required is very low and the precision is still acceptable. The details about this pitch extraction algorithm and experimental results could be found in [4].

In our system, speech feature contains MFCC, pitch and their first and second order differences.
3. NOISE AND CHANNEL COMPENSATION

Previous works on telephone speech recognition have shown that the main reason for this performance degradation is the variational mismatch caused by different telephone channels between the testing and training sets. Since non-stationary additive noise and non-linear distortion exists, the telephone network becomes a complex time-varying system. To compensate this mismatch, we use a Codebook-Dependent Dynamic channel Compensation (CDDC) algorithm developed in our group. This algorithm bases on maximum-likelihood estimation of telephone channel and could dynamically follow the variations inherent with the real communications network. CDDC could obviously improve the accuracy rate for telephony speech recognition. The details of this algorithm and its implementation on Mandarin telephony LVCSR system could be found in ICSLP2002.

4. OUT-OF-VOCABULARY REJECTION

Out-of-vocabulary (OOV) words are a common occurrence in almost every real speech recognition application and are a known source of performance degradation [1,2,3]. For any real-application-oriented system, the problem of in-vocabulary (IV) and out-of-vocabulary (OOV) classification remains a challenging one. It’s crucial to the robustness of system. Especially for many telephone-based applications, the purpose is to detect a certain set of keywords during continuous speech. The most common approach to do such keyword spotting tasks also depends upon the modeling of out-of-vocabulary speech events.

There are mainly two different types of OOV events exist in real applications:

- **OOV-I**: utterances containing mainly non-speech events such as laugh, cough, breathe, hesitation, clicks, smacks, line noise and distortion; these utterances are meaningless in the language.

- **OOV-II** utterances very similar in nature to the IV utterances. These utterances have their meanings in the language, but they are not contained in the system vocabulary.

For the first type of OOV, we could train following models to describe these noise utterances.

- **Non-speech Noise**: Laugh, cry, cough, sneeze, breath, hesitation;
- **Background Noise**: Keyboard clicks, slap, and background speech;
- **Line Noise**: Line distortion, buzz and impulsive noise;
- **Leading Speech Noise**: 
- **Trailing Speech Noise**: 
- **Other Noises**.

Using these models, we could construct a noise rejection network as illustrated in Figure 2. This better models the various noise sounds that may precede or follow the IV utterances and hence improve the recognition accuracy.

![Fig 2. Rejection strategy for OOV-I utterances](image)

Compared to the Rejection of OOV-I, the rejection of OOV-II is much more difficult due to its high comparability with IV sounds.

A conventional approach to deal with OOV-II is to process the searching score after the searching stage. One would expect a poorer score for speech not corresponding to an IV path than for a valid match to the path. A predetermined threshold applied on each path score, would enable discrimination between IV and OOV-II elements of the input speech. Since many of today’s speech recognition systems use word lattices or N-best-lists as a compact representation of a set of alternative hypothesis, score-based solution could be easily implemented on the word lattice or N-best list. This procedure is almost the same as the computation of confidence score from the word lattice. This solution could be simply depicted as follows Figure 3.

![Fig 3. Grammar network for n parallel vocabulary system](image)

Score-based rejection is performed by the two following tests:

\[
\text{reject if: } \begin{align*}
\text{Score}_{\text{best path}} & \leq t \quad \text{or} \\
\frac{\text{Score}_{\text{best path}}}{\text{Score}_{\text{next best path}}} & \leq t
\end{align*}
\]

Where \(\text{Score}_{\text{best path}}\) and \(\text{Score}_{\text{next best path}}\) are the best and next best path scores respectively and \(t\) is a threshold determined empirically.

This solution is efficient when the computation complex is of the most consideration. However, its effect is limited in the field.

The feature of our solution presented here is that the acceptance of IV utterances and the rejection of OOV-II utterances is implemented using a grammar that consists of two alternative parts, main grammar (consists of IV models) and rejection grammar (consists of OOV-II models). Clearly, it is not feasible to attempt to recognize all possible OOV-II spoken inputs. This method would be prohibitive in terms of the computation and storage requirements. A suitable compromise is to use generic speech models (sink or filler models) for OOV-II utterances.
Generic speech models are used in parallel with in-vocabulary models in a network to provide alternative hypotheses at a decision point. This solution is depicted in Figure 4. Here the entire word vocabulary is used in the search, where as the sink model is intended to cover OOV-II utterances.

The sink models are not intended to model any specific OOV-II sound. They are designed to fit well on all the OOV-II utterances. These sink models consist to build a rejection grammar which provides parallel searching paths for OOV-II events existing in the speech stream.

![Grammar Network Diagram](image)

Fig 4. Grammar network including $m$ parallel sink models and $n$ in vocabulary models in distinct parallel paths.

Input speech utterances are classified at the end point as follows:

$$
\text{reject if: } \frac{\text{Score}_{\text{vocabulary}}}{\text{Score}_{\text{sink}}} \leq t
$$

(2)

Where $\text{Score}_{\text{vocabulary}}$ is the best scoring IV path, $\text{Score}_{\text{sink}}$ is the best scoring sink path and $t$ is the rejection threshold.

There different methods to build this rejection network. We could especially collect speech utterances complementing the application vocabulary to train one or more generic speech models and use these complementing generic models to consist a rejection network for this purpose. Collecting complementing speech data for a specified application should be very careful. When the application domain changes frequently, it becomes a very exhausting work. Alternatively, we could use a superset of the main grammar. In this paper, we utilize the same acoustic models used in the main network to construct a superset of the main grammar as the OOV-II rejection network. So, collecting complementing speech data for each application is not needed. And this solution could largely facilitate the acceptance of IV and the rejection of OOV-II utterances.

We know that Mandarin Chinese is a monosyllable-structured tonal language. Unlike English, Chinese is character-based and words are not well-defined units. There are more than 10,000 frequently used characters. Clearly the computational overhead would be prohibitive for a rejection grammar, which constructed by all these characters. Each Chinese character’s pronunciation consists of an initial consonant and a final vowel. Tone also plays a major role in Chinese. Each character has one of five tones. There are only approximately 1,300 tone-specific syllables. We could use these syllables to construct the rejection grammar. In order to compress the computational overhead, we use non-tone-specific syllables to construct the rejection network. (There are approximately only 400 non-tone-specific syllables in Mandarin.) With appropriate path pruning, our searching algorithm could easily satisfy real-time requirements for telephone-based applications.

In this solution, the rejection grammar is a superset of the IV grammar, which covers all IV and OOV-II events. Through applying a penalty factor $p$ upon the rejection path, the IV/OOV-II classification could easily be done. The effectiveness of the sink path could be tuned with $p$. Our experiment shows that this super-generic speech model method is much better than score-based method. Additionally it is flexible between different application domains. And a syllable type rejection output makes it possible for later processing to identify the pronunciation (and possibly the spelling) of the OOV word.

5. EXPERIMENTS

It’s obvious that in order to maximize speech recognition performance, the training and testing sets should be as similar as possible. This means that we should use real telephone quality speech to train acoustic model for recognition. However there are still smaller amounts of real telephone speech available for training at present and develop dedicated speech database for telephone-based application is an exhausting work. In order to obtain telephone quality speech material for the acoustic model training, we utilize the high quality speech database developed for desktop LVCSR. We disposed this clean speech database with some methods including DSP down sampling, putting it through the real PSTN network and processing it by the GSM codec FR 06.10 (with and without channel error patterns). In preceding experiments, we compare the system performance using different training data. We find that the simulated speech data could also effectively improve the system performance obviously. In this system, we use totally about 230 hours Mandarin speech data for acoustical training.

We evaluated the system on two groups of telephony sets:

I. Type I test sets
   - Only contain IV and OOV-I data
   - **Stock**: Different isolated stock names, about 1000 utterances;
   - **Syl**: Different isolated syllables, about 6000 utterances;
   - **CDS**: Different continuous digital sequences, about 250 utterances.

II. Type II test sets
    - Contain IV, OOV-I & OOV-II data
    - **Reservation**: Continuous speech in the hotel reservation domain, more than 400 utterances.

First, we compare our OOV-I rejection solution with a simple energy-based detection method on type I test sets. The results are shown in Figure 5.
Our OOV-II rejection solution is evaluated on Reservation test set in terms of FRR and FAR, where FRR and FAR stand respectively for:

- **False Rejection Rate**: The ratio of incorrectly recognized and rejected IV utterances to all IV utterances.
- **False Acceptance Rate**: The ratio of accepted OOV-II utterances and incorrectly recognized IV utterances to all accepted utterances.

Obviously, our purpose is to decrease these two error rates as low as possible simultaneously. Tuning the penalty factor for OOV-II rejection paths, we could obtain a series of FRR and FAR.

Results are shown in Figure 6. In our system, we choose the \(p\) value at the EER (equal error rate, i.e. \(FRR = FAR\)) point.

Adding the OOV-II rejection grammar would inevitably degrade the system performance on test sets that only contain IV utterances. But we want the increase of WER for IV utterances to be as low as possible when applying rejection grammar.

In order to test the system under this condition, we do OOV-II rejection on CDS test set. The system detailed results are listed in Table 1.

![Fig 6. Tests for OOV-II rejection](image)

<table>
<thead>
<tr>
<th></th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Rejection</td>
<td>6.8</td>
<td>1.5</td>
<td>8.1</td>
<td>16.4</td>
</tr>
<tr>
<td>Energy Detection</td>
<td>6.2</td>
<td>1.7</td>
<td>2.4</td>
<td>10.3</td>
</tr>
<tr>
<td>OOV-I Rejection</td>
<td>6.1</td>
<td>1.3</td>
<td>1.3</td>
<td>8.7</td>
</tr>
<tr>
<td>OOV-I Rejection &amp; OOV-II Rejection</td>
<td>5.7</td>
<td>2.9</td>
<td>0.9</td>
<td>9.5</td>
</tr>
</tbody>
</table>

Tab 1. System summary percentages on CDS test set.

6. DISCUSSION

The results in figure (5) show that compared with no rejection conditions, our OOV-I solution decreases the WER about 14% and 11.6% for Stock and Syl test sets respectively. Unlike these two test sets, utterances in CDS are continuous speech streams. Utilizing OOV-I Rejection solution, WER even decreases about 47% on CDS set. Compared to energy-based detection solution, our OOV-I solution is a more reliable one. From the details of system performance listed in Table 1, we could find that the decrease of WER is mainly due to the decrease of insertion error.

Inevitably, utilizing the OOV-II grammar would decrease the system performance on IV utterances. But the test results listed in Table 1 show that this degradation is limited. When OOV-II grammar applied, the IV WER increases about 9%. But the robustness of the system is largely increased.

It should be pointed out that compared to real application environments, the utterances in our test sets have a relatively higher SNR. So the effectiveness of our OOV rejection strategy could not be shown entirely. In our on-line experiments, it could show an even more obvious progress.

7. CONCLUSION

In this paper, we first briefly present the framework of an application-oriented telephone speech recognition system and then we mainly describe the OOV rejection strategy implemented in our system. It is well known that OOV is a crucial problem to the robustness of every commercial recognizer. Through a series of benchmark tests, it is shown that our OOV rejection strategy could obviously improve the system performance in real application environments. It should be noted that our OOV solution only operates at the level of acoustic level and it is really domain-independent. So the transplant of this system into different application domains is very convenient. Further more, using a syllable-based rejection solution, arbitrary speech utterance could be decoded into a sequence of IV words and syllable units, which could be analyzed by subsequent domain-dependent processing blocks.

8. REFERENCES