AN OVERVIEW OF UYGHUR SPEECH RECOGNITION

Abstract: This paper reviews the progress of the Uyghur speech recognition research from the Uyghur acoustic features, speaker identification, keyword spotting, continuous speech recognition, etc. Firstly, this paper described the key technologies of the Uyghur speaker recognition from the feature extraction and transformation, the application of the hybrid model. Secondly, implement the Uyghur keyword spotting (KWS) by using the filler model, and tell the function of syllable lattice and confusion network. Finally, construct the Uyghur acoustic model and language model in continuous speech recognition based on continuous hidden Markov model. It also explores the applications of the Uyghur acoustics, prosody features in the post-processing.

Keywords: Speech Recognition; Acoustic Models; Language Models; Uyghur

1. Introduction

The automatic speech recognition (ASR) technology began in the 1950s. With the rapid development of computer hardware and software technology, and the gradual improvement of digital signal processing, machine learning, statistical learning theory, pattern recognition theory to lay the foundation for the research of speech recognition. In early 1970s, the speech signals linear predictive coding technology (LPC), dynamic time warping (DTW), vector quantization (VQ) was emerged. In 1980s, the hidden Markov model (HMM) based on the statistical theory replaced the method of traditional template matching. After that, there have been many successful applications based on the hybrid model of hidden Markov model with the Gaussian mixture model (GMM) or with the artificial neural network (ANN).

Uyghur belongs to the Turkic and Altaic, Uyghur is an agglutinative language, a wealth of affixes append to the same stem may constitute a very large scale vocabulary, and the lack of text and speech corpus, restricting the development of the Uyghur language speech recognition. The Uyghur speech recognition research started in the 1990s, with the establishment of the Uyghur speech acoustic parameters corpus, Uyghur speech recognition research has been made some progress. In 1994, we developed a Uyghur isolated-word speech recognition system with the recognition units of 1200 syllables, and the vocabulary of 40000 words [3]. Since 2000, the Uyghur continuous digital recognition system based on the method of DTW and VQ has been developed by Xinjiang Multi-lingual Information Key Laboratory. A HMM-based Uyghur continuous speech recognition system [4][6], a Uyghur telephony speaker recognition system based on a hybrid model of the Gaussian and support vector machine (SVM), and the Filler model-based Uyghur keyword spotting system have been born in our laboratory. According to the different tasks of speech recognition, this paper will illustrate the progress of the Uyghur speech recognition research from the Uyghur speaker identification, keyword detection and continuous speech recognition.

2. Acoustic Features of the Uyghur

Now China's official Uyghur is based on Arabic alphabet and Latin Uyghur as a supplement. Uyghur Phonemes include 8 vowels (V) and 24 consonants (C), and the Latin Uyghur is described in table 1.

Table 1: The Latin Uyghur

<table>
<thead>
<tr>
<th>V</th>
<th>a</th>
<th>E</th>
<th>i</th>
<th>O</th>
<th>u</th>
<th>ö</th>
<th>ü</th>
<th>ä</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>n</td>
<td>M</td>
<td>l</td>
<td>K</td>
<td>j</td>
<td>h</td>
<td>ğ</td>
<td>g</td>
</tr>
<tr>
<td>C</td>
<td>f</td>
<td>D</td>
<td>Ć</td>
<td>B</td>
<td>ž</td>
<td>z</td>
<td>y</td>
<td>x</td>
</tr>
</tbody>
</table>
Figure 1 show the vowel chart in type of Joos, the vertical axis is first formant, the horizontal axis is second formant and the ellipses represent vowels’ degree of dispersion. We completed this statistics with manually annotated continuous reading speech data.

![Vowel Chart](image)

Figure 1: Uyghur acoustic vowel chart

With the constitutive rules of Uyghur Syllable, each syllable has one and only one vowel, and one vowel can form a separate syllable. Uyghur vowel harmony based on articulation harmony, and rounded harmony is appended, so the former is the most serious part, and the latter is relatively unimportant.

In the word stem, if the vowel in the first syllable is a front (back) vowel, the vowel in the syllable after the first one is also a front (back) vowel. While the harmony phenomena in affixes part are often affected by the last syllable of the word stem, and does not appear with [o], [ö], [e]. On the effect of the allophone and the vowel reduction in unrounded front vowel, such as [i], [e], which can take both front and back vowel in the harmony of these vowels, and they are also impacted by the consonant before them. Uyghur is a kind of accented language, word accent is usually on the last syllable, when append affix to a word, the accent will be moved to the last. Consonant loss and vowel reduction phenomena in the Uyghur word lead to accent moving forward.

3. Uyghur Speaker Recognition

The speaker’s speech is able to demonstrate person's physiological characteristics. Speaker recognition can be divided into two tasks: speaker verification and speaker identification. According to the test speech, the former’s task is to determine whether the speaker and the alleged reference speaker are matched or not. The latter will use test speech to judge the speaker belong to which one of the multiple reference speakers. Speaker identification can be divided into an open set identification and a closed set identification.

3.1 Feature Extraction and Transformation

The acoustic features that can characterize the speaker’s personality characteristic mainly includes: the linear prediction coefficients and its derived parameters, which are consistent with the channel parameter model. For representing of speaker’s physiological differences (such as the excitation source, channel characteristics), speaker recognition also choose features from short-term spectrum, pitch contour, formant bandwidth and its trajectory, Robust parameter cepstral coefficients, the dynamic characteristics result from the difference of static characteristic, etc. Each parameter given above has its own merit on the characterization of speaker’s features, so feature fusion is required, but directly superimposed would increase the characteristic dimension and redundancy, and resulting in lower recognition performance. The transformation of the Uyghur language feature combines with principal component analysis (PCA) firstly, and then uses Fisher criterion for getting a set of high discrimination features. We also use the linear discriminant
analysis (LDA), linear heterogeneous discriminant (HLDA) analysis to implement the hybrid feature dimension reduction processing.

3.2 Models of Speaker Recognition
Speaker recognition models, generally divided into the probability of statistical models (such as GMM, HMM), and the decision model (such as SVM, ANN), the former reacts the similarity of similar data, the latter reacts differences of difference data.

SVM is commonly used to solve the problems which encountered in nonlinear, high dimension, and limited sample of the pattern recognition problem in practical applications. In speaker identification system’s training phase, regard each speaker as a class, extract each speaker's speech feature vectors as the input of the various types, after training, the results of which forms a multi-class SVM support vector and generates the classification function. In the identification stage, it is also extracted from the speech signal of the test feature vector sequence as the input of the SVM model. The SVM model classifies each vector, and then does statistics of the feature vector to assign which class of vector is the most. Finally, the model chooses the speaker corresponding to this class as the identification result.

Figure 2: Depiction of GMM-UBM approach

The Universal Background Model (UBM) is a large GMM trained to represent the speaker-independent distribution of features. In the GMM-UBM based system, if the speaker’s training data can cover all pronunciations, it could be use the speaker’s data to modeling. If the data cannot cover all cases, it should be use the set of other speakers’ data to training UBM and adapting. Figure 2 shows these steps in GMM-UBM modeling. In Uyghur speaker recognition, we combining the advantages of both approach, describe as in table 2 [7].

329
Table 2: Results of the Uyghur speaker recognition

<table>
<thead>
<tr>
<th>length of time</th>
<th>10s</th>
<th>20s</th>
<th>30s</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>58.3</td>
<td>81.2</td>
<td>90.8</td>
</tr>
<tr>
<td>GMM-UBM/SVM</td>
<td>63.5</td>
<td>92.1</td>
<td>93.4</td>
</tr>
</tbody>
</table>

4. Uyghur Keyword Spotting

Keywords spotting technology is divided into: the spotting method which based on Filler model or continuous speech recognition. According to the method of continuous speech recognition, the document retrieval and content detection technology based on the syllable lattice and confusion network has become the focus research of keywords spotting.

4.1 Filler Model

Filler model absorbs all outside of the keywords in languages (such as out of vocabulary words) and non-linguistic phenomena (such as noise, etc.). The search network includes filler, keyword, which can reduce the insert and delete errors in syllable recognition. In order to prevent keywords filler model engulfed by the keywords, the search structure often sets a certain degree of reward to the keywords models or punishment to filler models.

Table 2 shows the results of the Uyghur keyword spotting based on filler model with different sample size and number of keywords.

Table 3: Word accuracy and length of recognition time

<table>
<thead>
<tr>
<th>num of keywords</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample size</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>500</td>
</tr>
<tr>
<td>word acc (%)</td>
<td>97.0</td>
<td>93.5</td>
<td>91.0</td>
<td>89.6</td>
</tr>
<tr>
<td>average time</td>
<td>0.11s</td>
<td>0.54s</td>
<td>0.91s</td>
<td>1.61s</td>
</tr>
</tbody>
</table>

4.2 Syllable Lattice and Confusion Network

An approach of keywords spotting based on continuous speech recognition uses the decoder’s output to form syllables lattice or confusion network, and these are stored as text files. And then the KWS system selects the candidate keyword from the syllables lattice or confusion network by keyword search algorithm. At last, it verifies and outputs by using the confidence assessment. Such spotting methods, without repeated recognition of speech documents, it only needs to keyword search from the syllable lattices or confusion networks, so it is efficient and suitable for large vocabulary keywords spotting.

Speech has not only associated with spectrum of each acoustic unit segment, but also affected by a variety of suprasegmental factors. Suprasegmental factors refer to a level beyond the context of the local speech, such as accent, intonation, speed, pause, etc. Uyghur keyword spotting research tries to put the prosody features into syllables lattice or confuse network for the secondary decoder to improve the identification accuracy.

5. Continuous Speech Recognition

With the establishment of the Uyghur continuous speech corpus (telephone speech and reading speech), laid the foundation for Uyghur large vocabulary continuous speech recognition (LVCSR). Figure 1 shows the main steps of LVCSR system. The Uyghur continuous speech recognition research, mostly based on hidden Markov model, combined with the characteristics of Uyghur language. We have made a lot of work in the choice of recognition units, the creation of pronunciation dictionary, the design of question sets and the construction of language models.
5.1 Acoustic Models

At present, the mainstream speech recognition system generally uses statistical modeling frameworks based on Bayesian decision theory [1]. Uyghur contains more than 5000 commonly used syllables, so when use syllable as the unit of recognition, totally have 5000 monophone models, and the triphones will be extended to over ten million models, which will be lead to the sparseness of training data, resulting in parameter estimation insufficient. Uyghur continuous speech recognition generally uses the 34 phonemes (including silence and short pause) as the recognition units. Taking into account of the impact of context phoneme coarticulation, when Uyghur monophone is extended to triphone models, it would generate nearly 40,000 triphone models. It will also lead to a serious shortage of training data for each model (especially some models’ training with less occurrences or not) and reduce the training efficiency, and too much storage space occupied. The model parameters shared strategy can reduce the scale of parameters to be estimated.

Uyghur continuous speech recognition is generally used the mechanism of decision tree clustering to binding similar state. The strategy of decision tree provided a classification by doing a combination of the top-down data-driven method and the experts’ knowledge. For the splitting of decision tree, we need to combine the Uyghur acoustic and linguistic knowledge to build the problem sets. Uyghur questions are generally constructed in accordance with the vowel, consonant, the difference of articulation manner and articulation place and simple questions.

In the hybrid model of HMM/GMM [5], the Gaussian mixture models calculate the output probability of each model state in the HMM framework. Via the EM algorithm based on the probability of the GMM and maximum likelihood estimate (MLE) criteria to achieve the best analog on the probability distribution of each of the categories. With the increase of the Gaussian mixture component, the recognition rate there will be some improvement, but it also increase the reestimation of required parameters. So the increased number of Gaussian has associated with the size of the training corpus.

5.2 Statistical Language Models

The statistical language model provides the context and semantic information between words. As for Uyghur, it includes over 20,000 common roots and more than 300 affixes. They are formed an ultra-large-scale vocabulary through the combination. We lack speech corpus of the Uyghur language [2]: Firstly, a large number of balanced texts are difficult to obtain, cause the low coverage of commonly used words. Secondly, the field of the text corpus is limited. The size of the dictionary relates to the real system’s recognition effect. Generally speaking, the vocabulary of the practical recognition system betweens 5-10 orders of magnitude. The Uyghur statistical language
model to sort through four areas of the text corpus (the medical, news, magazines and novels), and do statistics of high-frequency appeared words, then make dictionaries with the vocabulary of 10-100 thousand words.

Because of the splice between the different classes of speech signal, the person is difficult to distinguish without context. The language model can improve the discrimination of the acoustic model. At present, more sophisticated statistical language model is the N-gram, the greater n value makes the higher precision and complexity of the language model. If the value of n is small, word sequence would make some certain words ignored, and result in some combinations cannot be rolled back. General choice of n = 3 to construct trigram language model.

Table 3 shows the recognition results of the Uyghur LVCSR, we use 6497 sentences for training and 90 sentences for testing with different number of the Gaussian mixture in 3-gram language models.

<table>
<thead>
<tr>
<th>mixture</th>
<th>18</th>
<th>24</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>word acc (%)</td>
<td>86.70</td>
<td>89.60</td>
<td>92.02</td>
<td>92.14</td>
</tr>
<tr>
<td>sent acc (%)</td>
<td>53.33</td>
<td>58.89</td>
<td>66.67</td>
<td>64.44</td>
</tr>
<tr>
<td>length of time</td>
<td>35s</td>
<td>38s</td>
<td>41s</td>
<td>42s</td>
</tr>
</tbody>
</table>

Language model training text size and its distribution has some limitations and one-sidedness, reasonable words context does not appear in the training text. By expanding the size of the training corpus, the low-frequency words still cannot get enough statistical properties of parameter probability estimation. Smoothing is to adjust the parameter values in the language model, the increase of low probability and zero probability, reducing high probability of parameter values. Prevent the emergence of zero probability to improve the accuracy of the language model. Commonly used smoothing techniques (such as the Good-Turning Katz, linear interpolation smoothing) are selected in Uyghur language model.

6. Conclusion

This paper described the status of Uyghur speech recognition development, although some progress has been made, compared with major domestic and foreign-language research also apart very far. In order to promote the development of the Uyghur speech recognition, we needs to combine acoustic, linguistic characteristics, analysis the factors restricting correct recognition rate, make the application of new modeling techniques, and study the research experience from the speech recognition technology of large multilingual.

Reference
