ACOUSTIC LEVEL ERROR ANALYSIS IN CONTINUOUS SPEECH RECOGNITION

LUO Chunhua, XU Mingxing, ZHENG Fang
Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science & Technology, Tsinghua University, Beijing, 100084
LuoCH@sp.cs.tsinghua.edu.cn

ABSTRACT
In this paper, we make a detailed analysis on the errors that may occur in a continuous speech recognition system, and define two sets of judge rules to perform the error analysis. Using these judge rules, we can efficiently find the most important factors that influence the performance of our speech recognition system and know how to improve it. The experimental results show that our judge rules have the ability to identify the types of errors in our system. They are also consistent with some conclusions drawn by other experiments.

Keywords: continuous speech recognition, judge rules, error analysis

1.INTRODUCTION
As we know that a traditional speech recognition system consists of the following parts: model training phase and recognition phase. The model-training phase includes feature extraction, model structure definition and appropriate parameter estimation. The task of a recognition phase is using a correctly searching algorithm, finding the boundary of recognition unit, and output the acoustic candidates according to the information provided by the acoustic model.

We hope that such a system can work as well as possible, but the real performance is not desirable. If we test the performance of our acoustical model cross the training set data independently, the correct recognition rate will be above 80%, but when we use the same acoustic model in the Chinese Dictation Machine (CDM) the correct recognition rate (acoustic level output) will only be about 70%. Why it decreases so much? In order to improve the performance of the CDM system, we must find the real reasons that introduce the recognition errors. Based on this motivation, it is also important for us to identify several types of errors.

1.1 Causation of recognition errors
In a continuous speech recognition system, there exist multiple causes that bring final recognition errors, for example, the weak acoustic model, the imprecise end-point detection, the imperfect state decoding and so on.

In our continuous speech recognition system, Easytalk 2000[1], in order to speed the system we use an end-point detection algorithm discussed in detail in the following section. The state-decoding algorithm based on the speech segment is a modified frame synchronous search [2](FSS) algorithm. Apparently, any mistakes in these phases will cause the final fatal recognition errors. In this paper, two different judge rules are proposed from different aspects to indicate the causation of recognition errors.

1.2 End-point detection
The end-point detection algorithm used in Easytalk 2000 is the merging-based syllable detection automation [3](MBSDA). It utilizes the short-term frame energy, zero crossing rate, and the pitch contour to merge one or several adjacent highly similar frames into merged similar speech segments which are regarded to belong to the same state of one syllable. There are three levels splitting points provided by the detection algorithm.

1. Level-0 splitting point
It stands for a silence boundary provided by the detection algorithm. All these points are surely correct.

2. Level-1 splitting point
It stands for the syllable unit boundary. The speech segment between two points of this type must contain one or more syllables if the segment is neither silence nor noise.
Sometimes we call these two splitting points as temporal splitting points for the comparison purpose with the splitting points provided by the acoustic searching.

3. Level-2 splitting point
All syllables’ boundaries must be selected from/among these level-2 splitting points. In other words, it must be at a level-2 splitting point when a state transition occurs between two syllables in the acoustic search processing.
The following figure illustrates the above concepts.

![Figure 1. Types of splitting points](image)

### 1.3 Modified frame synchronous search algorithms

The state decoding used in EasyTalk is a statistical knowledge based frame synchronous search (SKB-FSS) algorithm using the differential state dwell distribution (DSDD). The SKB-FSS assigns a possible state dwell range to State 0 according to the syllable number range given by the MBSDA algorithm. For any other State s, the possible state dwell range is calculated from the average state dwell of all through-going states and the DSDD information of the current state.

The following aspects will be covered in this paper: the implementation of the error analysis task, experimental results and some conclusions.

### 2. ERROR ANALYSIS

Let $o_W$ denote the output of the CDM (EasyTalk 2000), and $r_W$ denote the right sentence content of the input data (wave or cepstrum file). If $o_W$ is not completely equal to $r_W$, we can say that some errors occur. In this case, let $e_W$ denote the error output. Note that we only consider the best output with the highest matching likelihood score. In this paper we use $P_{XXX}(W_r)$ and $P_{XXX}(W_e)$ to denote the likelihood scores of $r_W$ and $e_W$ according to the XXX search algorithms respectively.

In our experiment, we use the following three searching methods.

1. (Method A) Perform an overall Viterbi search according to the silence boundaries extracted from the head of cepstrum file. As mentioned above, if the output is not correct, we can obtain two search path scores, $P_{cep}(W_r)$ and $P_{cep}(W_e)$ according to this search method.

2. (Method B) Perform an overall Viterbi search restricted by temporal splitting points provided by the splitting algorithm for the wave file. Similar to Method A, if the splitting algorithm dose not segment the wave file correctly, some errors will occur, therefore we can get $P_{wav}(W_r)$ and $P_{wav}(W_e)$ by using this search method.

3. (Method C) Perform a local search in some different cases, for example, whether using temporal splitting point, syllable range, state duration control strategy, etc. In EasyTalk 2000, we use the modified frame synchronous search (FSS) method to produce acoustic search results. In this case, let $P_{fss}(W_r)$ and $P_{fss}(W_e)$ denote the right and error recognition candidates respectively.

#### 2.1 Error type definition

When we use Method A to deal with the input data, we only use the silence boundaries, so we are confident that if error occurs, it might mainly comes from acoustic model errors. We define this error type as acoustic model error (AME).

In Method B, because it not only use the temporal splitting point, but also use the acoustic model (for calculating the likelihood probability for certain sentence output), it will introduce both AME and syllable splitting error named SSE.

In Method C, which in fact is a modified version of the Viterbi method, using wider information to perform the search, such as temporal splitting point, syllable range, state transition control strategy and so on. All of these will introduce some different types of errors. In order to distinguish these errors, three modifications of Method B are proposed according to the following demands respectively.

1. Using the syllable number range

   In the search process, if the syllable number range of one search path does not satisfy what we specify according to the splitting algorithm, the path will be discarded. But if the syllable range is not accurate, some underlying correct paths will be pruned irrationally and a recognition error is introduced. We call this type error syllable range error (SRE). In this case, we will obtain two matching scores, $P_{ran}(W_r)$ and $P_{ran}(W_e)$.

2. Using the level-2 splitting point

   In the search process of Easytalk 2000, if a state transition occurs, it must satisfy the condition that the current transition point is a level-2 splitting point. In such a controlled searching, the searching space will be reduced dramatically and the reasonable paths will be excluded. Thus the Level-
Two Splitting Error (TSE) type error is introduced. Again we will obtain two scores, $P_{\text{sec}}(W_r)$ and $P_{\text{sec}}(W_e)$.

3. Combining 1) and 2)

This means that the acoustic searching must satisfy the syllable range and the level-2 splitting point restriction simultaneously. For the reason discussed above, more recognition errors will be brought by more severe searching method. At this time, we use ASE (acoustic search error) to indicate the type of errors. Additional two scores, $P_{\text{sd}}(W_r)$ and $P_{\text{sd}}(W_e)$ will be used.

In order to identify the causes of speech recognition errors, two different sets of judge rules are proposed in the following section where the following six conditions are evaluated.

1. $(1) - (2)$
2. $(2) - (3)$
3. $(3) - (4)$
4. $(4) - (5)$
5. $(5) - (6)$
6. $(6) - (7)$

2.2 Judge rule set A

Rule 1: If $(1) < 0$, the error is of type AME.

When we calculate $P_{\text{cep}}(W_r)$ and $P_{\text{cep}}(W_e)$, we use the same Viterbi search method, so if error occurred, it must come from the acoustic model. Let us use $W_s$ to denote the searching result. If $W_s = W_r$, then $P_{\text{cep}}(W_r) \geq P_{\text{cep}}(W_e)$, in this case we can assume that the acoustic model result is correct, but if $W_s = W_e$, then $P_{\text{cep}}(W_r) < P_{\text{cep}}(W_e)$, it means that the acoustic model is not so good as we expected. There must be some errors. According to this principle, we can calculate the reliability of our acoustic model.

Rule 2: If $(1) * (2) < 0$, the error is of type SSE.

The difference between the calculation of $P_{\text{cep}}(W_r)$ and $P_{\text{cep}}(W_e)$ is that: the calculation of $P_{\text{cep}}(W_r)$ uses extra information of temporal splitting point provided by cutting algorithm. So does the calculation of $P_{\text{cep}}(W_e)$. If the temporal splitting point is correct, we will be sure that using this knowledge will improve the performance of our speech recognition system, at least not decrease the recognition performance, so if $P_{\text{cep}}(W_r) > P_{\text{cep}}(W_e)$, $P_{\text{cep}}(W_r)$ should be also bigger than $P_{\text{cep}}(W_e)$, hence we can use the value of $(1) * (2)$ to determine the syllable splitting error.

Rule 3: If $(3) * (4) < 0$, the error is of type ASE.

Comparing to the calculation of $P_{\text{sd}}(X)$, $X$ stands for $W_r$ or $W_e$, the state duration control strategy is used when calculating $P_{\text{sd}}(X)$. This strategy is the main method to control state duration and the transition in the frame synchronous search algorithm. So if our state duration control strategy is good, $(3)$ and $(4)$ should have the same increasing or decreasing trend. Hence $(3) * (4)$ can reflect the acoustic search error.

Rule 4: If $(2) * (5) < 0$, the error is of type SRE.

We use the syllable range information in our modified Viterbi search algorithm, so we can use a similar judge rule to determine this type of error.

2.3 Judge rule set B

Rule 1: If $(1) < 0$, the error is AME.

It is the same as the definition of AME in the judge rule set A.

Rule 2: If $(1) >= 0 && (2) < 0$, the error is SSE.

Comparing to the definition of SSE in judge rule set A, this definition is stricter. It’s related to the definition of AME. When $(1)$ is larger than zero, we can make sure that acoustic model is right, so if $(2)$ is smaller than zero, it indicate that the syllable splitting algorithm has some problem.

Rule 3: If $(1) >= 0 && (2) >= 0 && (4) >= 0 && (5) >= 0 && (6) >= 0 && (3) < 0$, the error is ASE.

This definition has closely relationship with searching method $c$ for that we just want to know the influence of state duration control strategy on the performance of a speech recognition system. So this type of error will occur only when other types do not occur.

Rule 4: If $(1) >= 0 && (2) >= 0 && (5) < 0$, the error is SRE.

If $(1)$ and $(2)$ are both larger than zero, we can say that the acoustic model and temporal splitting points is good, so we can decide the syllable range error depending on the value of $(5)$.

Rule 5: If $(1) >= 0 && (2) >= 0 && (6) < 0$, the error is TSE.

Similar to the definition of SRE, we can get the judge rule of TSE.

2.4 What’s the difference between the two sets of rules?

In Set A, one type of error may be recalculated in another. In other words, they are dependent. For example: in SRE, it uses not only the temporal splitting point, but also syllable range, so it will cover SSE error, our experiment certifies this point. But in Set B, the
errors are independent, so the sum of them will result in the total error.

The two sets of rules both can show the most influencing factor, which means that we can use the experiment result of judge rule set A to do qualitative research.

We think Set B is more reliable, because it judges more strictly, and the finally total error result are very approach our previous experiment result.

3.EXPERIMENT RESULTS

We use the mixed gaussian densities model as the acoustic model, in which one syllable unit model has six states and four Gaussian mixtures per state. The features are 34 dimensional MFCC and AR-MFCC.

We use M00, M02, M3, M04, M05, M06, M18, M20, M21, M22 as the train set for acoustic model training and M24, M25, M26 as the test set for performance testing. All of them belong to group A of 863 Database.

3.1 Using the judge rules set A

Table A.1 Training Set

<table>
<thead>
<tr>
<th>People</th>
<th>AME</th>
<th>SSE</th>
<th>ASE</th>
<th>SRE</th>
<th>TSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M00</td>
<td>0.182</td>
<td>0.060</td>
<td>0.530</td>
<td>0.198</td>
<td>0.079</td>
</tr>
<tr>
<td>M02</td>
<td>0.165</td>
<td>0.046</td>
<td>0.511</td>
<td>0.031</td>
<td>0.050</td>
</tr>
<tr>
<td>M03</td>
<td>0.298</td>
<td>0.067</td>
<td>0.536</td>
<td>0.065</td>
<td>0.083</td>
</tr>
<tr>
<td>M04</td>
<td>0.088</td>
<td>0.050</td>
<td>0.503</td>
<td>0.058</td>
<td>0.079</td>
</tr>
<tr>
<td>M05</td>
<td>0.098</td>
<td>0.054</td>
<td>0.545</td>
<td>0.061</td>
<td>0.065</td>
</tr>
<tr>
<td>M06</td>
<td>0.142</td>
<td>0.052</td>
<td>0.522</td>
<td>0.031</td>
<td>0.069</td>
</tr>
<tr>
<td>M20</td>
<td>0.226</td>
<td>0.069</td>
<td>0.468</td>
<td>0.274</td>
<td>0.052</td>
</tr>
<tr>
<td>M21</td>
<td>0.205</td>
<td>0.031</td>
<td>0.505</td>
<td>0.250</td>
<td>0.038</td>
</tr>
<tr>
<td>M22</td>
<td>0.225</td>
<td>0.046</td>
<td>0.511</td>
<td>0.192</td>
<td>0.069</td>
</tr>
</tbody>
</table>

Table A.2 Testing Set

<table>
<thead>
<tr>
<th>People</th>
<th>AME</th>
<th>SSE</th>
<th>ASE</th>
<th>SRE</th>
<th>TSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M24</td>
<td>0.324</td>
<td>0.094</td>
<td>0.474</td>
<td>0.131</td>
<td>0.046</td>
</tr>
<tr>
<td>M25</td>
<td>0.363</td>
<td>0.102</td>
<td>0.457</td>
<td>0.144</td>
<td>0.102</td>
</tr>
<tr>
<td>M26</td>
<td>0.257</td>
<td>0.086</td>
<td>0.407</td>
<td>0.232</td>
<td>0.090</td>
</tr>
</tbody>
</table>

3.2 Using the judge rule set B

Table B.1 Training Set

<table>
<thead>
<tr>
<th>People</th>
<th>AME</th>
<th>SSE</th>
<th>ASE</th>
<th>SRE</th>
<th>TSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M00</td>
<td>0.182</td>
<td>0.033</td>
<td>0.250</td>
<td>0.182</td>
<td>0.046</td>
</tr>
<tr>
<td>M02</td>
<td>0.165</td>
<td>0.035</td>
<td>0.167</td>
<td>0.021</td>
<td>0.035</td>
</tr>
<tr>
<td>M03</td>
<td>0.298</td>
<td>0.046</td>
<td>0.244</td>
<td>0.027</td>
<td>0.060</td>
</tr>
<tr>
<td>M04</td>
<td>0.088</td>
<td>0.044</td>
<td>0.219</td>
<td>0.056</td>
<td>0.060</td>
</tr>
<tr>
<td>M05</td>
<td>0.098</td>
<td>0.042</td>
<td>0.223</td>
<td>0.052</td>
<td>0.056</td>
</tr>
<tr>
<td>M06</td>
<td>0.142</td>
<td>0.042</td>
<td>0.211</td>
<td>0.027</td>
<td>0.056</td>
</tr>
<tr>
<td>M20</td>
<td>0.226</td>
<td>0.044</td>
<td>0.202</td>
<td>0.265</td>
<td>0.023</td>
</tr>
<tr>
<td>M21</td>
<td>0.205</td>
<td>0.023</td>
<td>0.246</td>
<td>0.234</td>
<td>0.017</td>
</tr>
<tr>
<td>M22</td>
<td>0.225</td>
<td>0.023</td>
<td>0.180</td>
<td>0.179</td>
<td>0.038</td>
</tr>
</tbody>
</table>

Table B.2 Testing Set

<table>
<thead>
<tr>
<th>People</th>
<th>AME</th>
<th>SSE</th>
<th>ASE</th>
<th>SRE</th>
<th>TSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M24</td>
<td>0.324</td>
<td>0.073</td>
<td>0.276</td>
<td>0.125</td>
<td>0.013</td>
</tr>
</tbody>
</table>

We also test the acoustic model performance solely. The syllable recognition error rate of the model is about 34%, which fits to the AME presented above.

4.CONCLUSIONS

The experiment results show that our judge rules have the abilities to identify the types of errors occurred in our CDM system. Based on the results, we can get the following important conclusions:

1.Apparently, we have a lot of work on improving the performance of acoustic model, for there are still more than 30% errors occurred.
2.The level-1 splitting point provided by the cutting process is believable, for it has 95% correct rate.
3.The level-2 splitting point provided by the cutting process is also credible.
4.The syllable range provided by the cutting process is varying for persons with different speech speeds, so it influences our search process much, and will inevitably decrease the performance of Chinese Dictation Machine. If we use a loose condition, what will occur? So more experiments are needed.
5.It is important for us to find a better state duration control (SDC) strategy for CDM, because there are above 50% errors introduced by SDC.

From the conclusions, we can find that the error analysis method proposed in this paper is very efficient, it can find out the key factors that affect the performance of the CSR systems and the aspects that we must, first of all, focus on. Further we must note that it is very easy for our error identification methods to be extended to a general situation.

5.REFERENCES