Investigation into Automatic Continuous Speech Recognition of Different Dialects of Northern Sotho

Hendrik J Oosthuizen, Madimetja A Mapeka, Madimetja J D Manamela

Department of Computer Science
University of Limpopo, P/Bag X1106, Sovenga, 0727, South Africa

Abstract

In this paper an automatic continuous speech speaker independent recognizer is constructed for the recognition of three main dialects of Northern Sotho. The recognizer is compared with a base-line system for Northern Sotho. The recognition accuracy was found to be less than that of the base-line system.

1. Introduction

Some progress has been made in construction of automatic continuous speech recognizers for the indigenous languages of South Africa. Speaker independent small to medium vocabulary recognizers for Northern Sotho and Setswana have been built at the University of Limpopo. In an ongoing project to increase the accuracy of recognition of these systems, various methods are being investigated.

In this paper the recognition accuracy of a hidden Markov model (HMM) system for three main dialects of Northern Sotho is contrasted to that obtained for a so-called base-line system. The latter did not take dialects into consideration. The Hidden Markov Model Toolkit [1] was used in the construction of the recognizer.

The dominating language in the Limpopo Province is Northern Sotho, with 51% of its population being mother tongue speakers of this language [2]. The Sepedi language serves as a standard dialect [3], with the following recognized dialect clusters:

- Central Sotho, covering the Pedi, Tau, Kone and Kopa tribes;
- Eastern Sotho, which accommodates the Kuretsi, Pai and Pulana tribes;
- North-Eastern Sotho which comprises the Phalaborwa, Lobjedu, Mamabolo, Letswalo, Mmametsa, Mahlo and Kgaga tribes;
- Northern Sotho, which covers the Mphahlele, Tshwene, Mathabatha, Maja, Mothapo, Matlala, Molepo, Tlokwa, Dikgale, Moletši and Hananwa tribes.

Three dialects which have strong ties with various other dialects, as indicated above, were selected for this study.

The Selobedu and Setlokwa dialects can be regarded as the major dialects which show the greatest dissimilarity with the main dialect, Sepedi. The regions where these dialects originated, are illustrated in figure 1 [4].

1. Geographical illustration of areas where data was collected. [4]

In most cases the variation of pronunciation in the different dialects is brought about by the position of a consonant or vowel in a word – this can be at the beginning of the word (prefix), the middle, and /or at the end (suffix). Examples of such differences are shown in table 1.

Table 1: Examples of phoneme occurrence in words.

<table>
<thead>
<tr>
<th>Phone occurrence</th>
<th>Locative suffix</th>
<th>Causative suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>/ng/</td>
<td>/s/</td>
</tr>
<tr>
<td>Dialects</td>
<td>English</td>
<td>Grass</td>
</tr>
<tr>
<td></td>
<td>Sepedi</td>
<td>Bjang</td>
</tr>
<tr>
<td></td>
<td>Setlokwa</td>
<td>Bjane</td>
</tr>
<tr>
<td></td>
<td>Selobedu</td>
<td>Bjanye</td>
</tr>
</tbody>
</table>

2. Speech data collection

The speech data was collected from volunteers whose home language was one of the dialects. This was done by preparing 150 different prompt sheets. The prompt sheets
Respondents were required to phone a Telkom toll-free number and to respond to voice prompts. The interactive system consisted of a Dialogic card in a Celeron PC connected to a basic rate ISDN-line. The software was designed and maintained by Molo Afrika Speech Technologies in Pretoria. The speech sampling rate was 8 KHz, 16 bit on a mono-channel.

3. Data validation and transcription

Since some of the respondents called from noisy environments, the recordings had to be carefully scrutinized to rid the data of extraneous noise. It was possible to obtain 30 responses from each dialectic region that were usable. A minimum number of responses per call of 15 out of the possible 21 (71%) was required for a call to be regarded as successful. The number of validated speech samples per dialect is given in table 5. The total number of validated utterances used in this project thus amounted to 1689 speech samples.

A large number of unwanted non-speech events was present in the recordings. These were cleared during transcription. The transcription was done at word level.

The way in which the speaker presents utterances can be associated with the differences in the pronunciation of words. This, in turn, leads to different representations of phonemes with respect to the dialects. The concept of dialects is thus significantly conveyed by the differences in specific phoneme pronunciation. The phoneme differences for the dialects under investigation are illustrated in table 6.

In order for the recognizer to function properly, a pronunciation lexicon has to be built. The lexicon in this
case covered the three dialects. Single words were thus mapped onto multiple pronunciations. As an example, the word bjang has multiple pronunciations and is represented as:

- **BJANG**  
- **BJANG (2)**  
- **BJANG (3)**

Hua and Schultz [5] employed this method and found it to be suitable. They tried to improve the recognition rate by using a triphone state tying tree. A marginal improvement was obtained. This method has also been implemented in the current study. Since in this case the system employs 43 phonemes, a total number of 129 decision trees should exist for the clustering method – three occurrences for each phoneme; at the beginning, the middle or at the end of a word.

### 4. Construction of the recognizer

As mentioned in the introduction, the HTK toolkit was used to construct the recognizer. The first step was the construction of the task grammar and dictionary. The main purpose of the task grammar is to define a closed set of words which the system will recognize. Any other words will be considered out-of-vocabulary (OOV) and will thus not be recognized. The task grammar was derived from the transcription files – individual words were extracted, sorted, and each word mapped to a unique pronunciation. This grammar had a unigram language model with all words equally likely. The different pronunciations of the same word were also equally likely. The original list contained a total of 842 unique words. Of these, 283 have a single pronunciation, 469 have two pronunciations, and 91 have three, according to the way the words are pronounced in the different dialects. The pronunciation lexicon thus contained 1464 words.

The speech signal was pre-emphasized by a factor of 0.97 and a hamming window of 25 ms was used with a 10 ms sampling rate. A 39 dimensional feature vector was used, consisting of 12 Mel-frequency Cepstral Coefficients (MFCC) with log energy, delta and acceleration coefficients added, all scaled around zero by subtracting the cepstral mean from all vectors. Three-state HHMs were trained with the training data and the HTK toolkit.

### 5. Results

A random selection of speech data from each dialect, balanced with respect to age groups, formed the test data. Thus 121 speech samples from each dialect were included in the test set, 363 utterances in total. The remainder formed the training data. No overlap between the training and test sets occurred. In order to compare the recognition rate of this recognizer with the case where a single dialect, Sepedi, was used, the data obtained by Modiba [6] were used. We will refer to this system as the baseline system in what follows. In the latter case the system was trained on 2706 sentences. For the test, 332 sentences comprising 1405 words were selected.

Evaluation tests for each dialect on its own were run, as well as for the mixed set of dialects. The results are indicated in table 7.

In the recognition results based on the test data, 72 (5.04%) OOV words occurred. With the speech sample used, it is not possible to eliminate the OOV words. If this could be achieved, the test results would improve significantly. Too many insertion errors occurred – further work need to be done in finding a suitable insertion penalty parameter value to balance the number of insertion and deletion errors. This should further improve the accuracy of the system.

#### Table 7: Results using pre-recorded test data.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepedi</td>
<td>W 76.97</td>
<td>595</td>
<td>132</td>
<td>5</td>
<td>139</td>
<td>53.61</td>
</tr>
<tr>
<td></td>
<td>S 36.36</td>
<td>121</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Setlokwa</td>
<td>W 79.19</td>
<td>609</td>
<td>138</td>
<td>7</td>
<td>172</td>
<td>47.95</td>
</tr>
<tr>
<td></td>
<td>S 40.50</td>
<td>121</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selobedu</td>
<td>W 72.02</td>
<td>604</td>
<td>167</td>
<td>2</td>
<td>278</td>
<td>25.99</td>
</tr>
<tr>
<td></td>
<td>S 10.74</td>
<td>121</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>W 75.06</td>
<td>1808</td>
<td>437</td>
<td>14</td>
<td>589</td>
<td>42.48</td>
</tr>
<tr>
<td></td>
<td>S 29.20</td>
<td>363</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where:  
- **W**: word  
- **S**: sentence  
- **Corr.** correct  
- **N**: total number of words/sentences  
- **S**: number of substitution errors  
- **D**: number of deletion errors  
- **I**: number of insertion errors  
- **Acc.**: accuracy

The system was also tested with live input. Twelve participants, six male and six female, four per dialect, were selected for this test. A very limited number of sentences were spoken by the participants, with the overall results indicated in table 8.

#### Table 8: Overall results for live input

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>52.7%</td>
</tr>
<tr>
<td>Sentence</td>
<td>16.7%</td>
</tr>
</tbody>
</table>

A comparison of the preceding results with the baseline system is depicted in table 9. As can be seen, the dialect recognition system was less accurate than the baseline system. The results, however, are promising and shows that with further refinement and a larger training corpus, a recognizer for continuous Northern Sotho speech
(including all dialects) can be constructed which may have an accuracy at least as good as the baseline system.

Table 9: Comparison of dialect and baseline systems.

<table>
<thead>
<tr>
<th></th>
<th>Pre-recorded speech</th>
<th>Live data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>word</td>
<td>Sentence</td>
</tr>
<tr>
<td>Dialect system</td>
<td>75.06</td>
<td>29.20</td>
</tr>
<tr>
<td>Baseline system</td>
<td>84.4</td>
<td>51.1</td>
</tr>
</tbody>
</table>

Various studies on recognition of dialects of European languages, as well as Japanese have been done. As an example, the report by the Speech Technology Group in Madrid, Spain, has reported on results of an investigation of Spanish dialects for a recognizer of natural numbers [7].

In future work, the system will be further improved by incorporating a proper language model, as well as experimentally adjusting the parameters to balance the number of insertion and deletion errors.

References