Automatic Speech Recognition of Spoken Proper Names

TI Modipa, HJ Oosthuizen., MJD Manamela
University of Limpopo
Department of Computer Science
Private Bag X 1106
Sovenga, 0727
Tel: (015) 268 3479
E-mail: thipem@ul.ac.za

Abstract—This paper covers the recognition of spoken Northern Sotho names focusing on first and second names. The speech recognizer was developed using sizeable collection of names of people with high frequency from the University of Limpopo database of registered students. The hidden Markov model toolkit was used to train the recognizer. The HTK toolkit uses phoneme recognition that indicates speech sound categories that are sufficient to differentiate between different words in a language.

Index Terms—hidden Markov model, spoken name recognition, automatic speech recognition

I. INTRODUCTION

THE recognition of proper names is an important and challenging component in speech technology. For Interactive Voice Recognition (IVR) system, it is important to retrieve the correct information for a specified name. It is important to have accurate pronunciations of proper names for speech recognition, speech synthesis, as well as for a speech synthesis component such as dialog systems, speech –to-speech machine translation, directory assistance, automated customer service and any state-of-the-art application that deals with natural language processing (NLP) [4]. Most names in Northern Sotho have multiple pronunciations, e.g. the name *mmabatho* may be written as *mabatho* but with the same pronunciation as a result of socio-linguistic phenomena [1]. In the second instance there is a single *m* but is pronounced as double *m*. Natural language processing tools such as named-entity recognizers and lexical disambiguation tools can benefit from such spelling correction [7]. Most of the Northern Sotho speakers use surnames including foreign origin or "borrowed" orthography. For instance, the surname *Chi*di has the pronunciation like *Tši*hidi, where the *ch* is pronounced as *Tši*.

Automatic speech recognition systems face a challenging task for the recognition of spoken names because most databases for spoken names are normally in the range of several hundred thousands [1]. This increases the complexity of recognizing all possible names due to lack of storage capacity.

Table 1 shows typical parameters used to characterize the capability of speech recognition systems. Those parameters that apply in this research are indicated in section three.

Table 1. Typical parameters used to characterize the capability of speech recognition systems [14].

Parameters	**Range**
Speaking mode	Isolated words to continuous speech
Speaking Style	Read speech to spontaneous speech
Enrollment	Speaker-dependent to speaker-independent
Vocabulary	Small (<20 words) to large (>20,000 words)
Language Model	Finite-state to context-sensitive
Perplexity	Small (< 10) to large (>100)
SNR	High (>30 dB) to low (< 10 dB)
Transducer	Voice-canceling microphone to telephone

Other difficulties of recognizing proper names as compared to ordinary words include, but are not limited to, the fact that patterns of stress in names are different depending on the language; the morphological structure of names is restricted in such a way that names cannot be broken down or decomposed like ordinary words. The pronunciation of names is a difficult process in itself and sometimes there is a contradiction in phonological patterns. A large number of different names exist within every class of speakers and as such it will require a large database to make a substantial coverage of those names [4].

An automatic speech recognition system can be developed with different recognition levels, namely, alphabet recognizer, phoneme recognizer, syllable recognizer, and word recognizer. The focus of this paper is on phoneme recognition. Fig. 1 shows an overview of the

speech recognition problem.

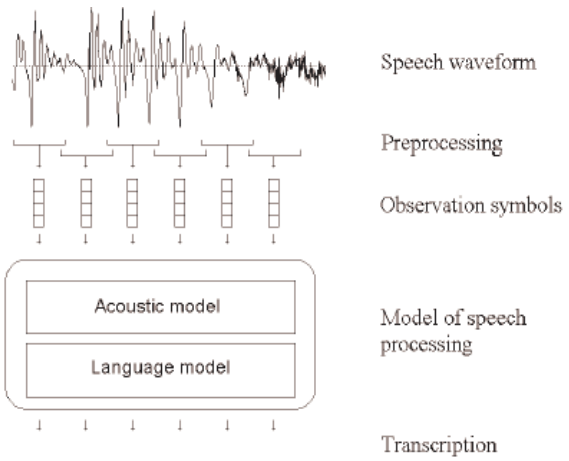

Fig. 1. Overview of the speech recognition problem

The second section of this research paper gives an overview of the automatic speech recognition approach. The experimental details, including the speech database, are described in section three. Preliminary results and discussions are described in section four. Section five gives the concluding remarks as well as an outline for future research.

II. SPEECH RECOGNITION

Automatic speech recognition is a process of converting a speech signal to a sequence of words. In general, a speech recognition system consists of the following components: signal processing, speech decoding, and adaptation. A speaker generates a word sequence which is passed through a communication channel to produce a waveform. The speech waveform is passed to the signal-processing component of the speech recognizer which will generate a parameterized acoustic signal. Using stochastic methods, the speech decoder component decodes the acoustic signal into a word sequence. Figure 2 shows a typical flow diagram of speech recognition system.

There are a number of automatic speech recognition applications that are currently in place, such as, voice dialing, call routing, simple data entry, and preparation of structured documents. These applications make use of a large proportion of spoken proper names.

The speech recognizer will determine what word string W was spoken given an input acoustic signal O. The acoustic signal is represented as a T-length string of spectral measurements $O = o_1,o_2,o_3, \ldots, o_T$ and W by a string of N words given by $W = w_1, w_2, \ldots, w_N$. The hypothesis $\widehat{W}$ is found by the maximum a posteriori recognizer as

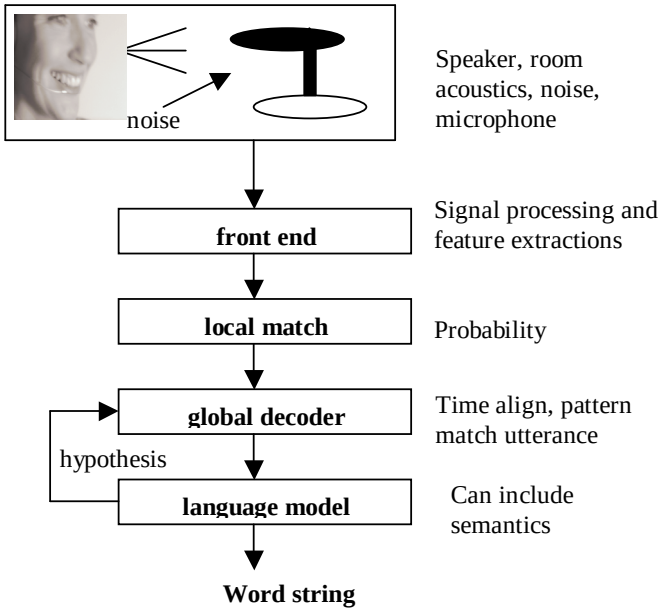

Fig. 2. Automatic Speech Recognition System [15].

$$\widehat{W} = \arg\max_{W \in \mathcal{W}} P(O \mid W) P(W) \tag{1}$$

where $\mathcal{W}$ represents the space of all possible word strings [12].

The output distribution of HMM state is modeled as a multiple Gaussian mixture model

$$P(o_t = o \mid s_t = s) = \sum_{j=1}^{K} \frac{w_{i,s,j}}{(2\pi)^{D/2} \left|\Sigma_{i,j,s}\right|^{1/2}} \exp\left\{(o - \mu_{i,s,j})^{\mathrm{T}} \Sigma_{i,j,s}^{-1} (o - \mu_{i,s,j})\right\} \tag{2}$$

where K is the number of Gaussian components, $w_{i,s,j}$, $u_{i,s,j}$ and $\Sigma_{i,s,j}$ are the mixture weight, mean and covariance matrix of the j^{th} component of the observation distribution of state s of the i^{th} word, T is the transpose and D is the dimension of the feature vector [12].

A. *Hidden Markov Model*

A Hidden Markov Model (HMM) is a statistical model which outputs a sequence of symbols or quantities. HMM is widely used because a speech signal could be viewed as a piece-wise stationary signal or a short-time stationary signal.

A hidden Markov model is a finite set of states each of which is associated with a probability distribution. The transitions among the states are governed by a set of probabilities called transition probabilities. The outcome or observation can be generated according to the associated symbol observation probability distribution. It is only outcome, not the state, which is visible to an external observer; states are hidden to the outside [13].

HMMs can also be trained automatically and are simple and computationally feasible to use. When dealing with large vocabulary continuous system there is a need for context dependency for the phonemes. In order to handle unseen contexts it would need tree clustering of the contexts.

B. Hidden Markov Model Toolkit

The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models. It has been chosen for this project because it has been in use worldwide for speech recognition research and some of the applications include research into speech synthesis, character recognition and DNA sequencing [16].

III. EXPERIMENTAL DETAILS

A. Telephone Speech database

A database consisting of first and last names (surnames) was created from the pool of current and former registered student records of University of Limpopo whose first language is Northern Sotho. From the database we extracted names occurring with high frequencies, as indicated in Table 2. And not all the names were considered for the training of the recognizer because the names were in the region of thousands.

Table 2. Partial list of names with their occurring frequency from University of Limpopo.

Name	**Frequency (%)**
Mphahlele	2.09
Mamabolo	2.03
Matome	1.96
Mokgadi	1.76
Malatji	1.39
Phuti	1.36
Letsoalo	1.32
Lesiba	1.21
Modiba	1.21
Matlala	1.15
Malesela	1.15
Tebogo	1.12
Chuene	1.09
Mpho	0.99
Mapula	0.99
Thabo	0.97
Tlou	0.96
Sello	0.91
Kwena	0.90
Masemola	0.82
Molepo	0.81
Moloto	0.78
Khomotso	0.76
Mogale	0.74
Kekana	0.74
Mogashoa	0.71

Prompt sheets containing randomized sets of three names were prepared to be used by recruited volunteers, all of them having Northern Sotho as first language. The respondents were required to use a toll-free telephone number to call and read in the information provided on the prompt sheet. Some of the telephone speech data that was recorded by the speakers was contaminated by background noise due to the fact that some were calling from busy streets where people and motor cars were passing.

Adobe audition software was used for the transcription of the telephone speech data collected from recruited participants. During the transcription, noise (non-speech utterance) that was captured during the recordings was removed in order to produce near-clean telephone speech data corpus.

B. Training the Speech Recognizer

Of the recruited participants we managed to get good recordings from 120 respondents, which generated 1839 sentences and 5404 words as indicated in Table 3. The recognizer has the following properties:

- It is speaker independent;
- Using limited vocabulary (names);
- Read speech;
- Trained for quiet environment;
- Telephone recorded speech;
- Continuous speech;

Table 3. Statistics of the training set for the native Northern Sotho ASR system.

Number of	**Training Set**	**Testing Set**
Speakers	80	40
Sentences	1226	613
Words	3603	1801

IV. PRELIMINARY RESULTS AND DISCUSSIONS

Two data sets were used for the training and testing of the recognizer. A pronunciation dictionary for the spoken names was also created prior to training the recognizer.

The recognizer accuracy was measured for spoken names which were grouped in a set of three names, indicating the

first name, middle name and last name. The recognizer generated good results especially with a single name as compared to three names, as indicated in Table 4. The reason is that with three names if one name is wrong the sentence will be judged wrong.

Table 4. Overall results generated by the recognizer.

Sentences % Correct	**Words** % Correct
45.85	62.85

V. Conclusion

The overall results of the system are good. The errors that occur are due to few utterances with phonemes such as *mpy, nkg, ntsh, sw, thw*. The system performance can be improved by utilizing more training data with sufficient coverage of all the phonemes. The results show that the automatic speech recognizer can operate under speaker-independent continuous conditions. This research project developed as a step towards building IVR for incorporation in more emerging market information and communication technology solutions as in e-service provision.

Future work will be aimed at developing a syllable-based recognizer and comparing it with the phone based recognizer for Northern Sotho names. We can further train our speech recognizer based on words using a different focus such as place names – towns, cities, provinces, etc.

Acknowledgment

We acknowledge the contribution of University of Limpopo Telkom Centre of Excellence for Speech Technology (TCoE4ST) and its funding partners – Telkom, Marpless, HP, and the NRF through THRIP.

References

[1] A. Sethy, S. Narayanan, and S. Parthasarthy, "A split lexicon approach for improved recognition of spoken names," Speech Communications, Vol. 48, No. 9, pp.1126–1136, 2006.

[2] M. Matrx, C. Schmandt, "Putting People First: Specify Proper Names in Speech Interfaces," Symposium on User Interface Software and Technology, California, 2-4 November, pp. 29–37, 1994.

[3] P. Louw, J. Roux, and E. Botha, "African Speech Technology (AST) telephone speech databases: Corpus design, contents and early validation statistics," [Online] Available: http://academic.sun.ac.za/su_clast/documents/EUROSPEECH2001.pdf

[4] A. F. Llitjos, "Improving pronunciation accuracy of proper names with language origin classes," Master Thesis, Carnegie Mellon University, August 2001.

[5] C. A. Kamm, C. R. Shamieh, and S. Singhal, "Speech recognition issues for directory assistance applications," *Speech Communication*, Vol. 17, No. 3-4, pp. 303–311, November 1995.

[6] F. Bechet, A. L. Gorin, J. H. Wright, and D.H. Tür, "Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How may I Help You?" *Speech Communication.*, Vol. 42, No. 2, pp. 207–225, February 2004.

[7] P. Ruch, R. Baud, A. Geissbühler, "Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record," *Artificial Intelligence in Medicine*, Vol. 29, pp. 169–184, 2003.

[8] A. Weber, and A. Cutler, "Lexical competition in non-native spoken-word recognition," *Journal of Memory and Language*, Vol. 50, No. 1, pp. 1–25, January 2004.

[9] J. Zhang, D. Shen, G. Zhou, J. Su, and C. Tan, "Enhancing HHM-based biomedical named entity recognition by studying special phenomenon," *Journal of Biomedical Informatics*, Vol. 37, No. 6, pp. 411–422, December 2004.

[10] N. Collier, and K. Takeuchi, "Comparison of character-level and part of speech features for name recognition in biomedical texts," *Journal of Biomedical Informatics*, Vol. 37, No. 6, pp. 423–435, December 2004.

[11] K. Yamamoto, T. Kudo, A. Konagaya, and Y. Matsumoto, "Use of morphological analysis in protein name recognition," Journal of Biomedical Informatics, Vol. 37, No. 6, pp. 471–482, December 2004.

[12] V. Venkataramani, S. Chakrabartty, W. Byrne, "*Gini*support vector machines for segmental minimum Bayes risk decoding of continuous speech," [Online] Available, www.sciencedirect.com.

[13] K. Wong, "Speaker adaptation with subspace regression classes," Master Thesis, Hong Kong University, August 2000.

[14] V. Zue, R. Cole, W. Ward, "Speech Recognition," [Online] Available, http://cslu.cse.ogi.edu/HLTsurvey/ch1node4.html.

[15] "Automatic Speech Recognition," http://www.informatik.uni-bremen.de/agki/www/ik98/prog/kursunterlagen/t2/node3.html

[16] "What is HTK," [Online] Available, http://htk.eng.cam.ac.uk/

Biography

T Modipa has completed his BSc (Hons) at University of the North. He is currently attached to the Telkom Centre of Excellence for Speech Technology at University of Limpopo and pursuing MSc. in Computer Science. His research interest is in automatic speech recognition.