Automatic recognition of spoken proper names with respect to Northern Sotho[1]

TI Modipa, Prof HJ Oosthuizen, MJD Manamela

thipem@ul.ac.za oosthuizenhj@ul.ac.za jonasm@ul.ac.za

(015)268 3479 (015)268 2169 (015)268 2627

Private Bag X 1106

Sovenga 0727

Abstract—This paper reports on the recognition of Northern Sotho first and last names by a trained HMM-based automatic speech recognizer (ASR). A comprehensive set of first and last names database was created and continuously read in by different participants to generate training and testing data. The recognizer thus developed will be used in voice-enabled tele-services where identification data such as names is supplied remotely by telephone.

I. INTRODUCTION

NAMES can be used as addressing mode in conversations to indicate a line of communication between people. They are regarded as part of everyone 's identity i.e. defining who we are. In Matt Marx [4] paper, it is pointed out that *names are rigid designators*, indicating that their referents do not or cannot change.

In order to initiate a conversation between two people, one party will pick up a phone and key in the phone numbers of the person they want to speak to. In this case the phone number is used as an addressing protocol for communication to take place between these people. An e-mail address can also be used as addressing protocol to point to the people to whom a conversation is intended. Names also carry the same weight in conversations, but names do not change when people move from one place to another as do other *flexible designators*, like phone numbers and e-mail addresses [4].

Ethnic origin of names plays a major role in the pronunciation of names [3]. There are more than four different dialects in Northern Sotho (Sesotho sa Leboa). Even if a name is in Sesotho sa Leboa, it may be pronounced differently from one area to another. This issue increases the challenges faced in name recognition as pointed out in III. Proper names are the most difficult ones for people to pronounce, the reason behind this being that most parents become creative and name their children in a way that would remind them of something else. This makes it difficult for human beings to pronounce such names properly.

[1] *This research is supported by grants from Telkom, HP, Marpless, and the NRF through THRIP*

II. SPEECH RECOGNITION

The interaction between computers and users has primarily been through the use of keyboard and mouse. In the past 30 years the spoken language interface has become more beneficial to physically challenged people. People can nowadays also interact with trained computer system in their mother tongue thereby reducing the cost of training individuals for the use of computer systems.

A spoken language system is divided into three categories: speech recognition system, which is the focus of this paper, text-to-speech system, and spoken language understanding. According to Ariadna Font Llitjós [2], speech recognition can be defined as the process of converting speech to text. This process whose main components are shown in figure 1 involves analyzing an audio signal to determine the words uttered by a speaker.

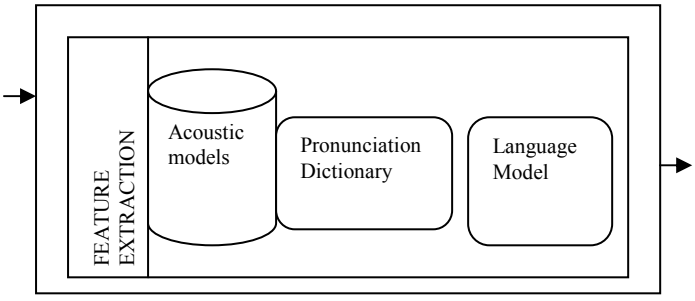

Figure 1. Components of an ASR System.

In order to determine the words spoken by the speaker; a speech recognition system does so by focusing on the phrases that are matched with the highest likelihood. In the process of speech recognition, the speech recognizer has a network of segments, each with ranked a list of phoneme labels [7].

Another part of the process of speech recognition process is to match the network of acoustic segments with their phoneme labels to the phoneme strings allowed by the grammar. The art of speech recognition includes techniques for conducting this complex search efficiently, so that speech recognition results can be returned without noticeable delay.

III. CHALLENGES IN NAME RECOGNITION

There is difficulty in recognizing proper names as compared to recognizing ordinary words.

- A large number of different names exist within every class of speakers. As such it will require a large database to make a substantial coverage of the names.
- Names emanate from different regions and areas, this results in a situation whereby one name can have more than one pronunciation.
- Pronunciation of names is a difficult process on its own. Sometimes there is a contradiction in phonological patterns.
- Morphological structure of names is restricted in a way that names cannot be broken down or decomposed like ordinary words.
- The patterns of stress in names are different depending on the language.

IV. HIDDEN MARKOV MODEL

Our approach to speech recognition uses a popularly used stochastic method based on the hidden Markov models (HMMs). A hidden Markov model is a finite set of states each of which is associated with a probability distribution. The transition among the states is 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.

At the beginning, the system will start with one of the states, according to the initial probabilities. Then at a regular time interval, a transition to another state will occur, according to the transition probabilities. When a state transition occurs, an observation is emitted. An emitted observation can be associated with the transition link, the source state or the destination state [1].

V. EXPERIMENTS

A large database of names is collected which comprises of the names originating amongst the Northern Sotho speaking population from around the country and mainly in the Limpopo province. Only the first and last names were considered for the study whereby non-native Christian names (usually English or Afrikaans names of Western origin) were excluded. Since the ASR system will be implemented for telephonic applications, Northern Sotho native speakers will be recruited to phone and/or read a list of names. A balanced group of respondents (more than 100) will be recruited for this purpose to build a speaker independent system. The hidden Markov model toolkit (HTK) will be used to train the recognizer with respect to Northern Sotho since it is the most spoken language in the province. HTK was adopted because it is primarily used for speech recognition research, speech synthesis, and character recognition. It uses words or phonemes as modeling units and mathematical processes estimate model parameters and in this study the recognition is at the phoneme level.

VI. CONCLUSION

In conclusion, the recognizer built in this study will used by native and non-native speakers of Northern Sotho. Since the system will be trained with speech data from a large number of speakers, it is expected to be very robust to recognize speech from any Northern Sotho speaker. This will be made possible by accommodating people from different dialects, a colleague's area of research and focus.

REFERENCES

[1] Kwok-Man Wong, "Speaker adaptation with subspace regression classes," Thesis, Hong Kong University, August 2000.

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

[3] Dr. M. Spiegel, "The difficulties with names, Overcoming barriers to personal voice services", http://www.speechtechmag.com/issues/8_3/cover/1993-1.html

[4] M. Marx, "Putting people first: Specifying proper names in speech interface", Marina del Rey, California, November 2 – 4, 1994.

[5] J. E. Hamaker, "Sparse Bayesian methods for continuous speech recognition, " Dissertation, Mississippi state University, September 2002.

[6] A. Venkataraman, "A statistical model for word discovery in transcribed speech," Association for Computational Linguistics, 2001.

[7] R. Kassel, "How speech recognition works," ScanSoft inc,http://www.microsoft.com/speech/docs/How_Speech_Works_Article.htm