

Speech Emotion Recognizer for Northern Sotho

Pressinah Moloto, Jonas Manamela, and Nalson Gasela

Telkom Centre of Excellence for Speech Technology

Department of Computer Science

University of Limpopo, Private Bag X1106, Sovenga 0727

Tel: +27 15 268 2798, Fax: +27 15 268 3487

email: pressinah@gmail.com; [jonas.manamela, nalson.gasela}@ul.ac.za](mailto:{jonas.manamela,nalson.gasela}@ul.ac.za)

Abstract—As a component of human-machine interaction, automatic speech recognition (ASR) technology allows a computer to identify words that a person speaks and converts them to written (or equivalent) text format. A speech emotion detector is very important for the proper handling of man-machine interaction and dialogue. It is therefore necessary to automate and identify a speaker's emotional state when using speech recognition tools. In a public service environment, detection of emotions in communication episodes is more likely to deliver correct and accurate assessment of the service rendering process. This paper proposes and describes the enhancement of a speech detection system to enable it to detect emotions in a Northern Sotho speech. The Hidden Markov Model (HTK) is used to capture and calculate the characteristics: pitch, mel-frequency cepstral coefficients (MFCCs) and formants of speech signal for emotion detection. We first investigated the ability of the people to detect emotions by a person. Preliminary results showed that 72% of the speech data recorded for 54 emotional speech utterances were manually correctly evaluated.

Index Terms—automatic speech recognition, emotional speech, Northern Sotho

I. INTRODUCTION

While speech technology research on systems that can recognize emotions is advancing, there has not been much work done on detection of emotional speech state using local African indigenous languages. Northern Sotho is one of the official languages of South Africa spoke by at least 9% of the total population of South Africa [9].

Human-computer interaction will be close to natural if computers can be able perceive and respond to non-verbal communication such as emotions. Emotional speech recognition can also be employed by psychologists to extract speech characteristic in real-time [2]. In emergency situation – which is the application focus on our research project, the envisaged ASR system will contribute towards exploring how emotional speech recognition could be used to improve e-service delivery in public services and corporate business environments. This may be done by improving the emergency responses, distinguishing the seriousness of emergency calls for more effective prioritization. The system may also help in providing feedback to an emergency call operator or supervisor for monitoring purposes.

The remainder of this paper is outlined hereafter. Section

II discusses the background of automatic speech recognition and emotional speech recognition. Section III discusses the design of the system, the contents of the database, speech emotional recognition and the analysis of the system. Sections IV, V and VI deal with conclusion, acknowledgements and references respectively.

II. BACKGROUND

Automatic speech recognition is a process through which a computer system can recognize the words spoken by a person and converts them to text [3]. Automatic speech recognition systems today find widespread application in tasks that require a human-machine interface, such as *automatic call processing* in telephony. It is widely accepted from psychological theory that human emotions can be classified into six emotions: *fear, surprise, anger, disgust, sadness and happiness* [4]. A study based on emotional utterances of fear, anger, happiness and a normal (unemotional) state was conducted to investigate how well humans and computers can detect emotions in spoken messages [6]. The speech emotional data was recorded and classified for training and testing. The study pointed out *pitch* as the main vocal cue for emotions. Other acoustic features that help in determining emotions include *vocal energy, frequency spectral features, formants and temporal features such as speaking rate and pausing*. In [7], the overall accuracy of the system in recognizing emotions was 63.5% and it was found that the computer ability to recognize emotions is at the same level as the recognition by humans.

III. EXPERIMENTAL APPROACH

A. Research Design

As an attempt to create this emotion recognition system, an HMM-based recognition toolkit, the Hidden Markov Model Toolkit (HTK) is used in our research project. The toolkit's flexibility, its ability to adapt easily to speaker's voice and speaking styles and its less memory requirements influenced its choice over other speech recognition toolkits such as Sphinx, Shout, Attila, SCARF etc. The selected speech processing toolkit is freely available at <http://htk.eng.cam.ac.uk/>.

B. Sampling and Data Collection / Database

Emotional speech data will be collected across various

relevant sources within the domain of emergency situations. In this research study, we shall also use *simulated or acted speech* - speech which is expressed in a professional and deliberated manner.

Emotion	Northern Sotho Sentences/Phrases	Evaluations
Normal	- Naa o bolela Sesotho? - Ga ke bolele Sesotho - O ile gae	66%
Anger	- Ke tla go bolaya - Ntlogele wena - Ke kwele go lekane	79%
Fear	- Thusa! - Ke a hwa! - Mollo!	71%

Table 1. Example sentences and the average results

A preliminary database has been created consisting of short Northern Sotho emotional utterances from different speakers. A total of 6 non-professional speakers (3 males, 3 females) were asked to record the sentences. A total of 54 emotional utterances were generated with 6 utterances for each emotional state. It may not be a surprise that anger and fear are emotions that are mostly found in emergency situations. For this reason, our database consists of utterances with *anger, fear* and *normal* speech state. 10 professional and non-professional Northern Sotho speakers evaluated the recordings and the average results are shown in the Table 1.

C. Emotion recognition

There is a variety of temporal and spectral features that can be extracted from human speech. However, it is argued that statistics related to pitch conveys considerable information about the emotional state of the speaker. To detect the emotion from speech signal, we use statistics relating to the *pitch, mel-frequency cepstral coefficients (MFCCs)* and the *formants of speech* as inputs. In order to capture these characteristics, the following statistics will also be calculated from the pitch: *mean, median, variance, the average energies of voiced and unvoiced speech* and the *speaking rate*.

D. Data Analysis

As the ASR system recognizes the emotions from the utterances, it must also determine the speech recognition accuracy. A few less literate to professional native Northern Sotho speakers will also help in evaluating the accuracy of the system in recognizing emotions. Finally, the results of computer-based detection of emotions and that produced by the recruited participants will be compared.

IV. CONCLUSION

In our research project, we explore three emotional states: *anger, fear* and *normal* state. While most people find it challenging to portray or recognize emotional content of a verbal speech, we plan to explore more on these data to see the behavior of the pitch and other speech characteristics before a larger database containing data from emergency situations can be used.

V. REFERENCES

- [1] Juang H.B. and Rabiner L.R. (2005): "Automatic Speech Recognition - A Brief History of the Technology", *Elsevier Encyclopaedia of Language and Linguistics, Second Edition*.
- [2] Ververidis D. and Kotropoulos C. (2006): "Emotional speech recognition: Resources, features, and methods", *Speech Communication*. Vol. 48(9), pp. 1162-118
- [3] Huang X, Acero A, and Hon H. (2001): "*Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*", Prentice Hall PTR,
- [4] Busso C. et al. (2004): "Analysis of Emotion Recognition Using Facial Expression, Speech and Multimodal Information", *Proceedings of the 6th International Conference on Multimodal interfaces*, pp. 205-211
- [5] Dellaert F, Polzin T, Waibel A. (1996): "Recognizing emotions in speech". *Proc. ICSLP 1996, Philadelphia, PA*, vol. 3, pp. 1970-1973
- [6] Petrushin V.A. (1998): "How well can people and computers recognize emotions in speech?" *Proc. of the AAAI Fall Symposium*, pp. 141-145.
- [7] Petrushin V.A. (2000): "Emotion recognition in speech signal: experimental study, development and application." *Proc. of the 6th International Conference on Spoken Language processing*, pp. 454-457.
- [8] Athanaselis T, Bakamidis S, Dologlu I, Cowie R, Douglas-Cowie E, Cox C. (2005): "ASR for emotional speech: clarifying the issues and enhancing performance". *Neural Networks*, vol. 18, pp. 437-444.
- [9] Census 2001: Census in brief. Pretoria: Statistics South Africa. (2003), [online]. Available: <http://www.statssa.gov.za/census01/html/CInBrief/CIB2001.pdf>, accessed on: 01 May 2012

Pressinah Moloto received her undergraduate degree in 2008 from the University of Limpopo and is presently studying towards her Master of Science degree at the same institution. Her research interests include Speech Technology and Software Engineering.