The Impact of Accents on Automatic Recognition of South African English Speech: A Preliminary Investigation

Audrey Mbogho
Department of Computer Science
University of Cape Town
Rondebosch, South Africa
+27-021-650-5108
Audrey.Mbogho@uct.ac.za

Michelle Katz
Department of Computer Science
University of Cape Town
Rondebosch, South Africa
+27-021-650-2663
michelleztak@gmail.com

ABSTRACT
The accent with which words are spoken can have a strong effect on the performance of a speech recognition system. In a multilingual country such as South Africa where English is not the first language of most citizens, the need to address this issue is critical when building speech-based systems. In this project we trained two sets of hidden Markov Models for isolated word English speech. The first set of models was trained with native English speakers and the second set was trained with non-native speakers from a representative sample of major South African accent groups. We compared the recognition accuracies of the two sets of models and found that the models trained with accented English performed better. This preliminary research indicates that there is merit to committing resources to the task of accented training.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing—Speech recognition and synthesis; H.5.2 [Information Interfaces and Presentation]: User Interfaces—Natural language.

General Terms
Languages, Human Factors

Keywords
Speech recognition, Accents, Hidden Markov Models.

1. INTRODUCTION
Computer scientists, and others, are interested in making it possible to interact with computers in much the same way we interact with one another, which is primarily through speech. Another motivating factor is the quest to make computers accessible to everyone. The standard mode of interaction, which is with the keyboard and mouse, excludes certain classes of people. Blind people, for example, cannot see icons in order to click on them. People who are unable to use hands due to ailments or injury cannot type and click easily, or at all. People with limited reading skills are similarly placed at a disadvantage by conventional modes of interaction, which rely on written text to identify icons since an image alone often cannot adequately indicate the purpose of an icon. While robust speech-based interaction could solve many of these problems, it is not easy to achieve due to a number of factors, including tone, pitch, volume, the age and gender of a speaker, the speed at which a word is spoken, background noise and the quality of the recording equipment—all these introduce uncertainty in the features used to distinguish an utterance. This paper addresses one particular challenge, namely, accent.

South Africa officially recognises 11 official languages: English, Afrikaans, isiNdebele, isiXhosa, isiZulu, Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga. There is a high variability in the pronunciation of English words especially in non-native English speakers. This suggests that any speech recognition software produced for South African users must be able to deal with a number of different South African accents. Ideally software produced for South African users should cater for all 11 languages and not just the pronunciation of English words by these speakers. This is a difficult problem because for South African languages there are acute deficiencies in our modelling abilities [6]. In this research, only the pronunciation of English words will be considered, and this focus is not without merit because of the special role English plays.

In South Africa, as well as several other African countries, English is often the primary language in which official information is exchanged. English is also widely used in these countries in ordinary, unofficial conversation. Even though, as we have stated, South Africa has eleven official languages, English is the de facto bridge that connects speakers of disparate languages, and, therefore, predominates. In 2001, of the approximate 44.8 million people in South Africa, 10.7 million (23.8%) spoke isiZulu as their home language. Thus, isiZulu is the most commonly used home language in the country followed by isiXhosa (7.9 million people or 17.6%) and Afrikaans (6 million people or 13.3%). Although English is only the 6th most common home language (3.6 million people or 8.2%), a significant proportion of the population has some competency in it. The African language speakers jointly make up 78% of the population [21].

The possibility exists to import speech systems that have been built in developed English speaking countries, such as the United States and the United Kingdom. Undoubtedly, such systems have been developed with the best resources available and are, therefore, as robust as the state of the art allows. However, IT solutions that are transferred unchanged from the developed world into the developing world tend to fail in the

SAICSIT '10, October 11–13, 2010, Bela Bela, South Africa.

new environment for a variety of application-specific reasons. For example, lack of bandwidth in Africa will adversely affect network-based technologies. One aim of this investigation is, therefore, to put speech systems through the "transfer test", and to address the weaknesses that this test reveals. In the context of the present work, the transfer test constitutes training a recogniser with native English speakers (this simulates the transplanted system) and testing it with a representative sample of speakers from a variety of accent groups in South Africa.

2. BACKGROUND

2.1 Isolated Word Recognition

The ultimate goal of speech recognition is to enable people to speak to machines with the same ease with which they speak to each other [11]. Speech recognition can be seen as a conversion of speech signals into symbols representing words [11], [13]. Recognition can be in terms of isolated words or continuous speech [2], [16]. Isolated words are significantly easier to recognize than continuous speech [1], and in this preliminary investigation we only address isolated word recognition. By limiting the problem to just a small set of words, it is possible to reduce the likelihood of error, because then there are fewer candidate outputs for each input. Such a limited vocabulary could nonetheless be useful. For example, the chosen set of words could be the commands of a user interface for an application.

2.2 Pronunciation

Words can be distinguished from each other by the way in which they are pronounced. Changing a sound slightly can change the meaning of a word completely. This idea is known as contrasts [20]. For example, the consonants "k" and "g" can be chosen to distinguish between "cane" and "gain". While the distinction between these sounds is significant in English, this is not necessarily the case for other languages. This can cause confusion with non-native speakers whose language perhaps regards the two sounds as interchangeable. While a human listener can easily use context to deduce the meaning of confused words, this problem is generally intractable for machines.

Sounds used contrastively, or phonemes, can be produced by using different parts of the vocal tract. For example using lips, teeth, tongue (in terms of the tip, blade, body or root) or nasal cavity affects the sound produced. Different languages may rely more heavily on a particular part of the vocal tract in comparison to other languages. For example Afrikaans is more of a guttural (using the pharynx wall or the larynx) language than English. This may affect how an Afrikaans speaker pronounces English words. Variant pronunciations, or allophones, of contrastively used sounds are also possible. The differences in allophones can be very slight and difficult to recognize when looking only at a speech signal. It is due to allophones that the sounds in the words "pan", tan" and "can" differ from "span", "Stan" and "scan". The sounds "p", "t" and "k" are called plosives and can be accompanied by a puff of air or aspiration as in "pan", "tan" and "can". However, the aspiration is not present in words beginning with "s".

Another phenomenon which the above example highlights is that of coarticulation. The articulation of one sound may have an uncontrollable effect on the articulation of neighbouring sounds. Once again, this is very difficult to pick up by looking at a speech signal both manually or by a speech recognizer. There can be patterns of stressed and unstressed syllables. This is known as rhythm. Intonation is closely linked to these two concepts and refers to the change in pitch of the voice throughout an utterance. Stress, rhythm and intonation together are called prosody. The usage of stress, rhythm and intonation varies from one language to the next. Some languages use different tunes to differentiate between individual words; these are called tone languages. This is not the case for English, but isiZulu and isiXhosa are tonal languages [6]. This means that speakers of English whose native tongue is isiZulu or isiXhosa may introduce tone to English, making automatic recognition more uncertain.

2.3 Hidden Markov Models

HMMs [16] have been used extensively for speech recognition. In isolated word recognition, each word is modelled as an HMM. Training involves using several speech samples of a given word to build a model for that word. These samples come from different speakers in order to achieve speaker independence, and in this study where we are addressing accents, these speakers are from different accent groups. In the recognition phase an unknown utterance is fed through each model in order to compute the probability that the word corresponds to the model. That is, $p(word|HMM_i)$ is computed for each HMM, and the HMM that gives the highest probability is chosen. Test utterances are put through this recognition procedure in order to assess the recognition accuracy of the models. We use HTK (the Hidden Markov Model Toolkit) from Cambridge University [23] for model training and testing. Although there are other toolkits available, HTK seems to be the more popular choice for research and applications in speech recognition [3, 4, 7, 9, 14].

3. RELATED WORK

3.1 Speech Variability

Different pronunciations will have different phonetic representation and therefore speakers with different dialects are likely to find speech recognition systems unusable due to high error rates [13]. Pronunciation was an important issue in [22] as the system was being used to teach English as a second language. In this situation, the system must include alternate, acceptable pronunciations. There has been much research for Chinese dialects and tones in Asian languages [4, 10]. Efficient speech recognition must be able to cope with different accents and speaking styles [15]. HTK, for example, uses pronunciation dictionaries which hold more than one pronunciation associated with each word [7]. IDEA (International Dialects of English Archive) was created in 1997 and holds an archive of downloadable recording of many dialects of English, including South African English. These recordings were found to be inappropriate for the research covered in this paper as the speech data was limited and of continuous speech. The AST (African Speech Technology) project has also created a speech database of South African accents [18]. A study of accents and the pronunciations in indigenous South African languages was completed in [24]. Although there have been many similar studies of South African languages in speech recognition, there is little evidence of research on the effects of accents on speech recognition accuracy.

3.2 Performance

Performance in speech recognition is measured in terms of accuracy and speed. Speech recognition performance has not yet reached a stage where it can be considered on the same level as human performance as machine error rates are more than an order of magnitude greater than human error rates [11,

12]. Due to low performance of automatic speech recognition systems it is difficult to extend their application to new fields [8]. Speaker variation is one of the major sources of error and can cause computational complexity to increase substantially [7]. Some believe that most improvements in performance in speech recognition are due to advances in microelectronics and are independent of work in speech technology [11]. However, although error rates are considered too high for many applications, development of speech recognition systems is still an ongoing process. It is possible for parameters to be "pruned" in the decoding stage in order to decrease computation time with moderate degradation of the word error rate [7]. In order to improve performance in speech recognition, a paradigm shift may be necessary from signal transcription to message comprehension, where the system not only recognizes words, but their meaning in context [11].

Performance related to isolated words and command recognition is significantly greater than that of continuous speech recognition [1].This increase in performance for isolated words is due to a number of factors. For example, there is no need to worry about word boundary locations in isolated word recognition [23]. For continuous speech, word boundaries are a problem as these boundaries may be blurred [1]. Word recognition with a limited vocabulary has simpler grammars and requires less speech data. In addition the words have equal probabilities of being spoken [19]. This paper will look specifically at isolated words, as recognition rates are higher than those of continuous speech and yet there are many applications which can be built around isolated word recognition.

3.3 Applications

In order for applications based on speech recognition to be possible, it is necessary for the recognizer to provide a measurable benefit to users, to be user-friendly and accurate and to respond in real time. Speech recognition can have applications in office/business systems, manufacture, telecommunications and healthcare [17].

The use of speech recognition technology in healthcare was studied in [5]. It was found that speech applications were used for dictation, speech-based interactive systems, speech controlled equipment and language interpretation systems.

A speech recognizer was used for computer aided instruction in order to teach English as a second language [22]. The system helped students to practice and improve their spoken English.

Speech recognition has been used to convert telephone speech to text [7, 21]. However, transcription of conversational telephone speech is a highly challenging task in speech recognition and state-of-the-art systems incur word error rates of 30%-40% [7].

It is possible to use speech recognition to summarize voice mail messages to text. In this case the system is also required to identify the important information in a message [9].

Table 1. Words used in training and testing

BREAD BUS CELLPHONE CHILDREN CLINIC EIGHT FIVE FOUR FRIDAY GARAGE GOODBYE HELLO HOSPITAL HOUSE MILK MONDAY NINE ONE SATURDAY SCHOOL SEVEN SHOP SIX SUNDAY SUPERMARKET TAXI THREE THURSDAY TRAIN TUESDAY TWO UNIVERSITY WATER WEDNESDAY WORK ZERO

4. EXPERIMENTS

In this study we refer to a set of models trained using utterances by native English speakers as the *English* model. We refer to the set of models trained using utterances by speakers from all accents used in the study as the *accented* model. Furthermore, to simplify the discussion, we will speak of *a model* when in fact we mean *a set of models* since each word has its own model. We wish to compare the performance of an English model with that of an accented model.

The question researched in this paper is, "How well can an accented model, trained with South African accents, improve recognition accuracy of words spoken with South African accents?" In other words, this experiment will test whether accuracy level for speech recognition in South African accents can be improved by using South African accents to train the recognizer. Training is a costly exercise both in terms of money and time. Therefore, it is important to establish whether it is worthwhile to invest in the increased amount of training required to cover all the major accent groups. Whereas we identified only 5 accents due to our study's limited scope, in reality there are many more, and implementing this in practice would imply a significant investment.

In order to address the research question stated above, it was necessary to obtain speech signals from speakers of different South African accent groups. Databases with such speech signals already exist, for example the AST project [18]. However this experiment is on isolated words and not continuous speech. Although the AST database contains a vast amount of data it is more appropriate for an experiment in continuous speech. Thus, we decided that creating a new speech database would be necessary. In our new collection each speech signal represents one of 36 isolated words shown in Table 1. The 36 words include digits, days of the week, basic foods and transport, i.e., words commonly found in every-day speech . The isolated word is sandwiched between two "silence" signals. "Silence" in this case means that the speaker is not speaking at that moment. However, there may still be background noises or static.

Five broad accent groups were identified for this investigation, namely: Afrikaans, African, Cape Coloured, English, and Indian. These are the same accent groups used by the AST project [18]. There are obvious limitations in this, or any, grouping of accents. One is that each group contains multiple subgroups. Another is that finding people whose speech typifies a particular group's way of speaking is difficult. Yet another is that how people classify themselves can be subjective, with some placing themselves in the group they prefer.

4.1 Experimental Environment

A standard microphone with a sensitivity of -58dB ± 3dB and frequency range of 50 – 13000Hz was used to collect the speech signals. The speech signals were collected from 50[1] speakers, most being university students and the breakdown according to gender is shown in Table 2.

Table 2. Breakdown of contributors by accent and gender

Accent	Males	Females
English	6	4
African Languages	6	4
Afrikaans	9	1
Indian	5	5
Cape Colored	3	7

As mentioned above, the speakers were mostly university students. This will have an impact on the results as there is an implied level of education and thus it is likely that the speakers' accents may not be as strong as they would be for the general public. The speakers being university students, also implies a certain age of the speakers; between 18 and 25 years. Each speaker provided 36 speech signals, one for each of the 36 words. The words were shuffled and chosen at random. The word was then recorded. This was necessary in order to ensure that the order in which the words were spoken had a minimal effect on the quality of the speech signal. For example, many users started off shy and the first recordings were softer and unsure. The speakers gained confidence towards the middle of the recording process but became bored and fatigued towards the end.

An ethical problem that was identified was that of not offending speakers in terms of their accents. Speakers often don't fall into one specific accent group and they may take offence when their accent group is wrongly (or sometimes even correctly) inferred. For this reason speakers were asked to identify their own accent groups. This however presented another problem where speakers put themselves in accent groups which were unexpected. For example, an Afrikaans speaker who went to an English school and spoke with an English accent (to our ears) identified his accent as Afrikaans. For this study, we left out speech samples that were clearly misclassified.

Two recognizers were created for comparison. The first, the English model, was trained using only English accents. The second, the accented model, was trained with all five accent groups. 20 speakers were used, each contributing one utterance of each word. The English model was trained with the utterances of 10 speakers, all of whom belonged to the English accent group. The accented model was trained with 2 speakers from each of the five accent groups. Thus the size of the training data was 360 utterances for the English model and 360 utterances for the accented model.

The test data was a separate set of utterances from 8 speakers, 2 from each of the 4 accent groups, excluding English. Thus the size of the test data was 288 utterances. The models were then tested with this data set in order to compare the accuracy of their recognition results. Accuracy was measured according to the word recognition rate, which is simply the percentage of words that were correctly recognised.

[1] Only samples from 28 out of 50 participants were used in the experiments in this paper. The full database is available at http://shenzi.cs.uct.ac.za/~honsproj/cgi-bin/view/2009/katz_mathai_sobey.zip/Speech_Katz_Mathai_Sobey/DownloadSpeechSA.html

4.2 Implementation

The entire process from data collection to testing the recognizer was accomplished on a HP dual core, with 2.40GHz processing speed and 4GB memory. A Linux operating system was used (Ubuntu, Jaunty Jackelope). The speech models are created using HMMs in HTK. A summary of the process is shown in Figure 1 where the rectangular shapes indicate steps in the process. The oval shapes indicate the input and output of the steps. The following method was used to create all three models. However, the training and testing data differed for each model.

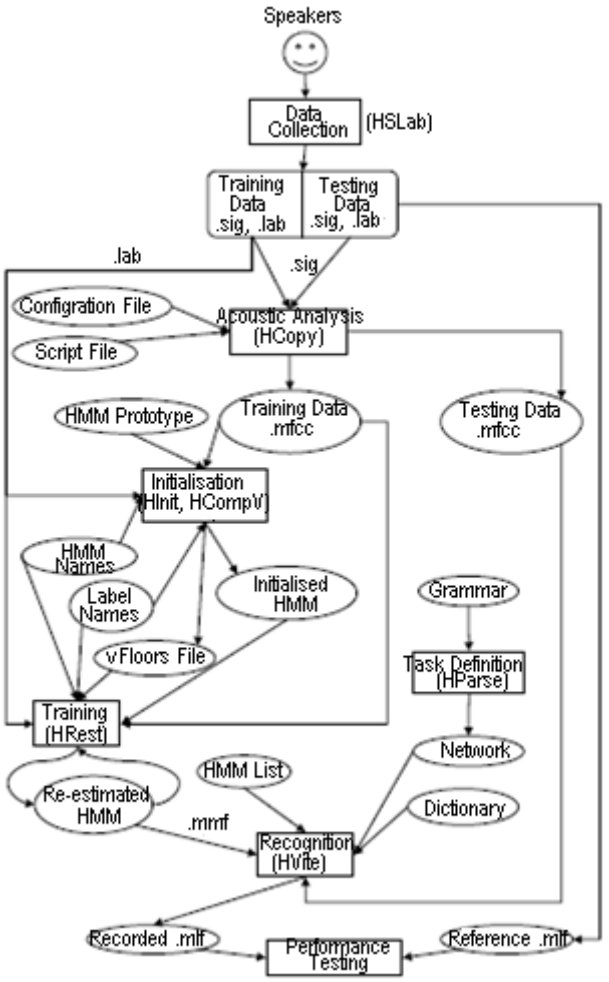

Figure 1. Training and Testing Overview

The HSLab tool in HTK was used to collect and label the speech signals. Labelling of speech signals involves identifying the starting and ending point of each word or phoneme in the signal. For the purpose of this experiment full words were labelled. The silence areas in a signal were also identified. Although this is an extremely time consuming process, the files must be hand-labelled as this cannot be automated (it would require a fully functional speech recognition system).

Acoustic analysis was performed using the tools provided within HTK. The signals were then broken up into overlapping frames every 20ms using a window size of 25ms and a target rate of 10ms. The number of MFCC coefficients was set to 12,

the number of filterbank channels was set to 26 and the length of the cepstral filtering was set to 22. The pre-emphasis coefficient was set to 0.97. For each signal frame 39 coefficient vectors were extracted. These parameters were as recommended in [23].

Each of the 36 words was modelled using a single HMM. In addition, an HMM was defined for silence or "sil".

5. TESTING AND RESULTS

Each model was tested for each speaker (not included in the training process) for the four accent groups: Afrikaans, African languages, Cape Coloured and Indian. Although it is possible to use the HResults facility in HTK for this process, more detailed results can be obtained using Microsoft Excel to compare reference and recorded files. The word to be recognized and the results from the recognizer were written to a comma separated values (csv) file. This was done for all 36 words. This step was repeated for each speaker for both models. The two words (the actual and recognized words) were compared and an average accuracy or recognition rate was found for each speaker and for each word in an accent group. The results were then analyzed in Minitab 15.

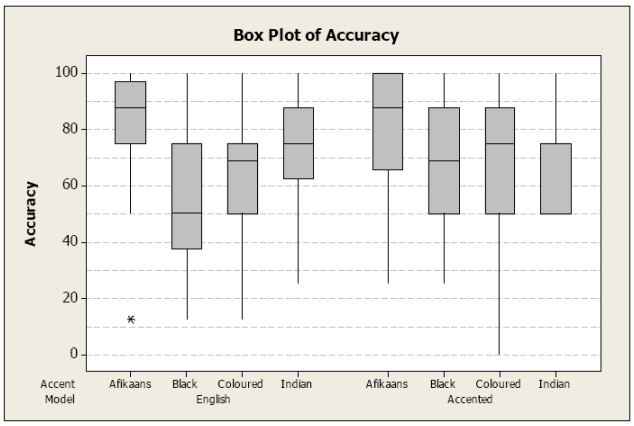

Figure 2. Recognition accuracy box plot

A box plot is shown in Figure 2. The diagram shows 8 box plots, one for each of the 4 accent groups for the 2 models: English and accented. The Afrikaans box plot for the English model indicates the lowest interquartile range and therefore the lowest amount of variability. However this box plot also indicates an outlier where the recognition rate was 37.5%. This outlier was identified as the recognition rate of the word "five". This outlier can be removed before doing any further testing in order to improve the quality of the results. The Cape Coloured box plot also had an increase in the size of the interquartile range (from 25% to 37.5%), suggesting an increase in variability. The medians for both Afrikaans (87.5%) and Indian (75%) box plots in the English model are equal to those of the accented model. However, the Black language and Cape Coloured box plots show increases in the medians from the English model to accented model. The Black language median increased from 50% to 68.75% while the Cape Coloured median increased from 68.75% to 75%. The whiskers for all box plots extend up to 100%. The Cape Coloured box plot in the accented model had the lowest reaching whisker at 25% (Interestingly this is lower than the outlier identified in the Afrikaans box plot for the English model).

Figures 3 and 4 show the histograms for observations in the English model and the accented model respectively. These histograms show that the accented model has a mean of 70.92% recognition rate which is slightly higher than that of the English model at 68.23%. The standard deviation also slightly decreased from 23.23 in the English model to 20.15 in the accented model. Both models can be fit to a bell curve, suggesting that they are normally distributed. A test for normality would be required to confirm this, but this goes beyond the scope of this paper, as we only wish to establish an indication of improvement in recognition accuracy.

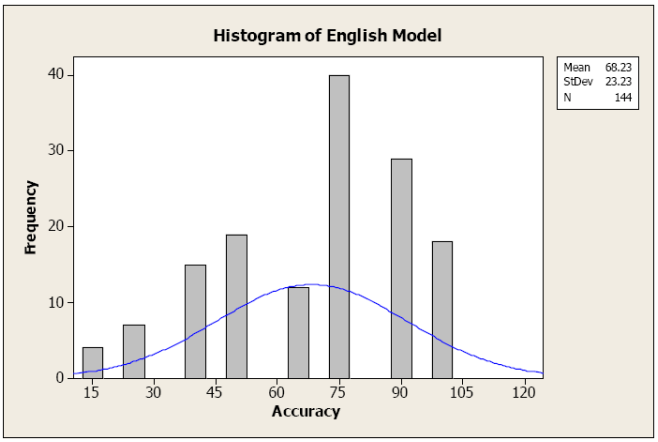

Figure 3. Histogram of English Model

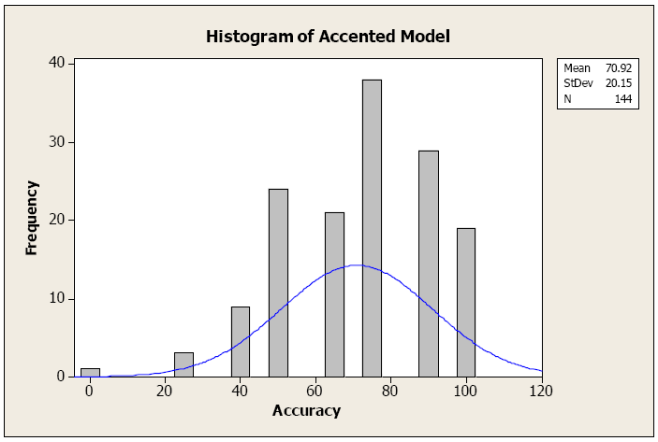

Figure 4. Histogram of Accented Model

In general, the English model showed a recognition rate of 68%, whereas the five-accent model showed a recognition rate of 71%, a 4.4% improvement. However, the diagram also suggests an increase in variability (i.e. decrease in consistency) in the accented model.

6. CONCLUSION

In this study, we created a database of isolated English words spoken in five major South African accents. We trained a set of models using native English speakers and a second set using speakers from five accent groups: English, Afrikaans, African, Cape Coloured and Indian. We found that recognition rates of accented test utterances improved when using the accented model versus the English model. Further analysis of the data indicated that more training data would be required to confirm these results. It is expected that with more training data, consistency in the data and recognition accuracy should increase. It is entirely possible that the disparities between South African accents are too great to lump them together in one accented model. This suggests another approach, which we are investigating, namely to train individual accent models, so that there would be a model for each accent. These would be coupled with an accent classifier. The accent classifier would

intercept a speech input, identify its accent and then channel it to the appropriate accent-specific recogniser. Our ultimate goal is to build a usable speech-enabled user interface that is robust to accents. This work could easily be extended to other parts of the world where regional accents exist.

7. REFERENCES

[1] Atal, B. 1995. Speech Technology in 2001: New Research Directions. In Proceedings of the National Academy of Sciences of the United States of America. 92, 22, 10046-10051.

[2] Bahl, L., Jelinek, F., and Mercer, R. 1983. A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-5, 2, 179-190.

[3] Cai, J., Bouselmi, G., Laprie, Y., and Haton J. 2009. Efficient likelihood evaluation and dynamic Gaussian selection for HMM-based speech recognition. Computer Speech and Language. 23, 147-164.

[4] Chen, J., and Jang, J. 2008. TRUES: Tone recognition using extended segments. ACM Trans. Asian Lang. Inform. Process. 7, 3, Article 10 (August).

[5] Durling, S., and Lumsden, J. 2008. Speech Recognition use in Healthcare Applications. In Proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia (Linz Austria, 2008). 473-478.

[6] Govender, N., Barnard, E., and Davel, M. 2007. Pitch Modelling for the Nguni Languages. South African Computer Journal. 38, 28-39.

[7] Huang, X. 1992. Minimizing Speaker Variation Effects for Speaker-Independent Speech Recognition. In Proceeding of the Workshop on Speech and Natural Language (Harriman New York, 1992). 191-196.

[8] Jeong, M., and Lee, G. 2008. Improving Speech Recognition and Understanding Using Error-Corrective Re-ranking. ACM Trans. Asian Lang. Inform. Process. 7, 1, Article 2 (February).

[9] Koumpis, K., and Renals, S. 2005. Automatic Summarization of Voicemail Messages Using Lexical and Prosodic Features. ACM Transactions on Speech and Language Processing. 2, 1, Article 1 (February).

[10] Lee, T., Lau, W., Wong, Y., and Ching, P. 2002. Using Tone Information in Cantonese Continuous Speech Recognition. ACM Trans. Asian Lang. Inform. Process. 1, 83–102.

[11] Levinson, S. 1995. Speech Recognition Technology: A Critique. In Proceedings of the National Academy of Sciences of the United States of America. 92, 22, 9953-9955.

[12] Lippman, R. 1997. Speech Recognition by Machines and Humans. Speech Communication. 22, 1-15.

[13] Markhoul, J., and Schwartz, R. 1995. State of the Art in Continuous Speech Recognition. In Proceedings of the National Academy of Sciences of the United States of America. 92, 22, 9956-9963.

[14] Morales, N., Toledano, D., Hansen, J., and Garrido, J. 2009. Feature Compensation Techniques for ASR on Band-Limited Speech. IEEE Transaction on Audio, Speech and Language Processing. 17, 4, 758-774.

[15] Mosur, R. 1996. Efficient Algorithms for Speech Recognition. PhD thesis, Carnegie Mellon University, May 1996. CMU-CS-96-143.

[16] Rabiner, L. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In Proceedings of the IEEE. 77, 2, 257–286.

[17] Rabiner, L., and Juang, B. 1993. Fundamentals of Speech Recognition. Prentice Hall.

[18] Roux, J., Botha, E., and du Preez, J. 2000. Developing a Multilingual Telephone Based Information System in African Languages. Second International Language Resources and Evaluation Conference. (Athens Greece, 2000).

[19] Smit, W., and Barnard, E. 2009. Continuous Speech Recognition with Sparse Coding. Computer Speech and Language. 23, 200-219.

[20] Spencer, A. 1996. Phonology: Theory and Description. Blackwell Publishers: Great Britain.

[21] Van der Merwe, I., Van der Merwe J. 2006. Linguistic Atlas of South Africa: Language in Space and Time. Sun Press: Stellenbosch

[22] Xie, H., Andreae, P., Zhang, M., and Warren, P. 2004. Learning Models for English Speech Recognition. In Proceedings of Conferences in Research and Practice in Information Technology (Dunedin New Zealand, 2004). 26, 323-329.

[23] Young, S., Evermann, G., Gales, M., Hain, T., et al. 2009. The HTK Book. Cambridge University Engineering Department: Cambridge

[24] Zerbian, S and Barnard, E. 2008. Phonetics of Intonation in South African Bantu languages. Southern African Linguistics and Applied Language Studies. 26(2), 235–254.