A Review for Transcribing Broadcast News for Low-Resource Languages

Kristen Jodie Basson
bsskri003@myuct.ac.za
Department of Computer Science
University of Cape Town
South Africa

ABSTRACT

This paper is a review of speech-to-text transcription of broadcast news for low-resource languages to create more documentation in the language. The nine low-resource South African Bantu languages have very few electronic textual resources. Thus, creating text resources using audio from a controlled audio environment would be beneficial for further research in areas such as search engines and machine learning. Reviewing acoustic models and the various toolkits available and which seems to perform the best for low-resource languages. As well as using the correct training approach to train the acoustic model is important in achieving the best result. Segmentation then plays an important role in reducing computation time, identifying speaker turns and identities as well as identifying and removing non-speech segments. Segmentation is also used to detect a change in speakers known as speaker diarization. Furthermore, data augmentation is used to achieve further accuracy in a model by decreasing the word error rate. Language models were evaluated using the Lwazi corpus. Phoneme and word language models for the South African languages were compared using perplexity, phoneme error rate and word error rate. Additionally, pronunciation dictionaries were automatically created for isiXhosa, Sesotho and Setswana using the NCHLT corpus and results from the languages were compared. There are many challenges to creating textual resources for the low-resource languages in South Africa, yet there are studies that provide a helpful starting point and guide to overcome these challenges.

1. Introduction

There are nine ethnic South African languages, many of which are drastically under-resourced, especially in terms of text-based electronic resources such as documents or books [4], [9]. Thus, the transcription of the South African Bantu languages would be beneficial to create more text documentation in those languages. This could further support experiments in areas such as search engines and machine learning. The easiest way to produce quality text documentation in these languages would be to use broadcast news. The South African Broadcast Corporation (SABC) broadcasts the news in all eleven South African languages. In these broadcasts, formal language is spoken and therefore elements such as code-switching and code-mixing can be eliminated, creating a controlled environment to capture audio [7]. This controlled environment can be described as recordings in a quiet studio setting and can be classified as baseline broadcast speech.

All automatic speech recognition (ASR) systems work in the same manner: training of the model and speech decoding. The models that can be used to facilitate audio transcription of broadcast news are acoustic models [14]. This literature review will discuss ASR technology and acoustic modelling as well as various acoustic model toolkits and training methods. An ASR system uses acoustic models which contain statistical methods. Toolkits can be used which make use of certain acoustic models. Various toolkits will be discussed in terms of low-resource languages and compared to find one that achieved the most accurate results. Accuracy is furthermore determined by using audio partitioning [11]. These segmentation methods reduce computation time and have other advantages. An example includes speaker diarization which is when a change in speakers is detected. Data augmentation is also another way to increase a system's performance yielding more accurate results [10]. Augmentation methods will be evaluated by considering experiments done by Jimerson et al. [14] where word error rate (WER) decreased due to augmentation. Language model performance will be evaluated using South African languages and their training data [13]. Lastly, different ways to create a pronunciation dictionary will be addressed along with examples of pronunciation dictionaries from different studies, including an example done on some South African languages for a pronunciation dictionary.
This review will provide clarity towards transcribing low-resourced languages and how that can be applied to the transcription of SABC broadcast news in the nine ethnic South African languages.

2. ASR and acoustic models

ASR systems convert speech into a text format. There are many components to an ASR system, one of which includes acoustic modelling. Acoustic models contain statistical methods used to model specific audio signals to their phonetic sounds. There are various ASR toolkits and frameworks that use different acoustic models [14]. Some examples are the Hidden Markov Model toolkit (HTK), Kaldi toolkit, JANUS speech recognition toolkit (JRTK), long short-term memory (LSTM) recurrent neural network (RNN) provided by the Persephone toolkit, and other deep neural networks

(DNN) frameworks. Depending on what is expected of your ASR system, different acoustic models would achieve individual results. Many studies deduce that there is a significant increase in performance when using DNNs for a system with limited resources [10]. This is further motivated by Zhao and Zhang [22] who reviewed the hybrid DNN/HMM ASR system. The results confirmed that the DNN/HMM-based system had better performance than an end-to-end model for under-resourced languages. As well as Nassif et al. [17] who reviewed neural networks as an acoustic model and found that hybrid models for DNN give better results than other models. Arisoy et al. [2] also found that Phoneme Error Rates (PER) were significantly less when a DNN acoustic model was used. On the contrary, when Jimerson et al. [14] used a DNN with a minimal amount of data from the endangered language of Seneca (155 minutes), WER was so high that the output was unusable whereas when using Kaldi, the WER was about 35% lower. Naidoo and Tsoeu [16] tested English and isiXhosa with three open-source toolkits: Kaldi, HTK and CMU Sphinx. The overall results were the best for the Kaldi toolkit, although specifically for isiXhosa the CMU Sphinx toolkit performed the best.

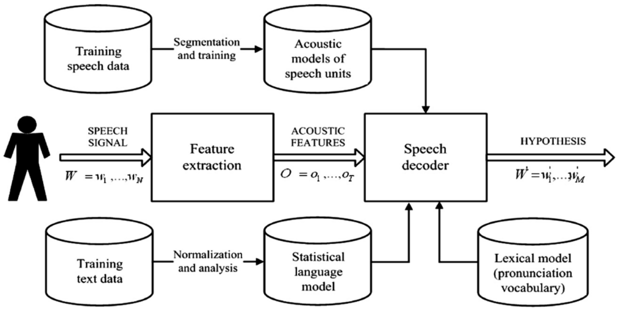
Training speech data
Segmentation and training
Acoustic models of speech units
SPEECH SIGNAL
$W = w_1, \ldots, w_N$
Feature extraction
ACOUSTIC FEATURES
$O = o_1, \ldots, o_T$
Speech decoder
HYPOTHESIS
$W' = w'_1, \ldots, w'_M$
Training text data
Normalization and analysis
Statistical language model
Lexical model (pronunciation vocabulary)

Figure 1. Architecture of a state-of-the-art automatic speech recognition system and its components. [4]

These acoustic models were trained using large amounts of collected speech and text data. There are various machine-learning approaches to training datasets, relevant to audio transcription, such as: supervised, unsupervised and semi-supervised learning [17]. Supervised learning is a type of machine learning using labelled data and training the datasets by giving them the correct input and desired output. Unsupervised learning uses input data with unlabelled output data to train the learning algorithm. Semi-supervised learning is used with a large amount of input data where some of the data is labelled and others not. Since under-resourced languages, especially South African Bantu languages, lack text documentation for training data, an unsupervised or semi-supervised method is recommended [4]. It was not until recently that many South African languages had no data available to develop an ASR system [13], [3]. The AST and Lwazi corpora were first developed to solve this problem. Since then, the national centre for human language technology (NCHLT) is the most recent corpus that was developed and contains more than 50 hours of speech in each of the South African languages which can be used in ASR and associated systems.

3. Segmentation

Audio partitioning is used to further divide acoustic signals into homogeneous segments in terms of a speaker, gender, and bandwidth [8], [15]. The content is then labelled and structured and non-speech segments are identified and removed. Acoustic segmentation is done using the Bayesian information criterion (BIC) framework [6]. Labels for segments can be classified as F0-baseline broadcast speech, F2-spontaneous broadcast speech, F3-speech over telephone channels, F4-speech with background music, F5-speech from non-native speakers and other segmentations are labelled FX [7], [6]. The results of Brugnara et al. [6] showed that the F0 and F3 produced the lowest WER. This is due to the conditions being less noisy and having higher audio quality. There are multiple advantages to using audio partitioning [11]. Firstly, valuable information can be extracted such as speaker turns and speaker identities. Secondly, speech discontinuity at speaker turns can be avoided. Lastly, identifying and removing non-speech segments as well as dividing the data into smaller segments simplifies decoding and significantly reduces computation time. Gauvain et al. [11] concludes from experiments that their transcription quality was good considering it was for speech over background music. This is because it is usually ensured that the background music level is low enough for the speaker to be heard clearly. However, the sections that were the most difficult to transcribe and with the highest error rates were speech from non-native speakers and overlapping speech that occurred in interviews or voice-overs for translated sections. Furthermore, it is emphasized that model accuracy is greatly dependent on the size of the audio and textual corpora. The larger the corpora to work from the greater the accuracy.

Segmentation in terms of speakers helps identify speakers and their turns [18]. A change in speakers is also known as speaker diarization (SD). For example, Deléglise et al. [8] used a segmentation system which clustered by speaker allowing the system to produce non-overlapping speech segments which correspond to speaker turns. This happens frequently in broadcast news where a change in speakers occurs due to the change in news sections. For example, a switch from headline news to weather broadcasts. Besacier et al. [4]; Patel et al. [18]; Deléglise et al. [8] all used speaker diarization. Patel et al. [18] explains that diarization is a module that partitions input audio recordings to the change in speakers. Furthermore, the LIUM diarization toolkit was used to determine the change in speakers and was in conclusion successful.

A language called Cree is a low-resource language that has hardly-any text resources available [12]. Radio broadcast archives presented 1200 hours of audio data and 260 000 words of text data. Two segmentations were used: speaker-independent, which is a system trained to respond to a word regardless of who is speaking and speaker-dependent, which refers to a system that is trained to a user's individual speech pattern, dialect, or language. Unfortunately, for the speaker-independent scenario, the WER was too high for accurate transcription. However, the speaker-dependent scenario had a WER of 24.6% and a phoneme error rate (PER) of 8.7%. These error rates are low enough to significantly increase transcription speed. This shows that segmenting data as speaker-dependent could potentially allow for lower phoneme error rates as well as word error rates. Using segmentation in an acoustic model is valuable in breaking up the parts needed to be transcribed and will in turn yield a better overall performance of the system.

4. Data augmentation

Data augmentation is another way to increase a system's performance yielding more accurate results [10]. It is a technique used to increase the training dataset and is a necessary and efficient method to utilise with limited data resources. Gales et al. [10] use data and transcription generation where either audio data is artificially generated, or additional transcriptions are generated using a semi-supervised approach. The two forms of data augmentation that were explored were vocal tract length perturbation (VTLP); and semi-supervised training. It was deduced by the results that the best overall performing system used only VTLP data augmentation. Moreover, Jimerson et al. [14] used data augmentation for the endangered language of Seneca. Spontaneous conversations of the endangered language that were recorded and transcribed were used to train an acoustic model. These recordings were quite diverse since they had been made over many years with varying recording equipment. The acoustic augmentation techniques that were used were: noise addition, speed augmentation and pitch augmentation. For the augmentation results to be observable, the original unaugmented sample was kept then a speech sample was augmented 15 times and another 25 times. Particularly these sample sizes were 155 minutes, 4500 minutes, and 7500 minutes. The results for the acoustic augmentation showed a significant decrease from the original data's WER to the WER for the data augmented 25 times. An alternative form of augmentation is cross-language augmentation is used in transfer learning [21]. This is where models and resources used in one language are transferred and used in the other language. Additionally, this method was used on the 155 minutes of original data with an English speech model which substantially reduced the WER [14]. Even though both transfer learning and data augmentation indicated significant reductions in WER for a DNN, the Kaldi toolkit trained on only 155 minutes yielded a lower WER overall. Text augmentation was also used to create synthetic sentences from the original 1843 sentences in the corpus. Unfortunately, for the text augmentation, there was an increase in WER.

5. Language models

Perplexity measures how well the language model performs, the lower the perplexity the better. Henselmans et al. [13] did a study on the South African languages using the Lwazi corpus and some of their results show the perplexities of phoneme and word language models. For the phoneme language model, their results show that English and Afrikaans have a higher unigram perplexity than the other nine South African Bantu languages. This suggests that the nine ethnic languages' phoneme sequences are more predictable. The explanation for this is that these languages use open syllables allowing phoneme sequences to be more predictable. Additionally, the bigram phoneme perplexities were even lower for the Southern Bantu languages. Furthermore, the PER corresponds to the perplexities where English and Afrikaans have a higher PER than the other nine languages. Noticeably, isiZulu seems to perform much worse than the other languages despite having a low perplexity and a large amount of training data in comparison to the other languages. According to Henselman et al. [13] this result was unforeseen considering the similarities between isiZulu and isiXhosa.

The word language model, unfortunately, recorded high unigram perplexities for isiZulu, isiXhosa, isiNdebele and Siswati despite the large size of their training datasets. This trend seems to continue in bigram perplexities although noticeably less. The nine ethnic South African language perplexities decrease substantially, especially isiXhosa which is only in the bigram perplexities considered average in comparison to the other languages. Additionally, due to high perplexities, it is observed that the WER is thus also high. Since repeated phrases were also removed from the training datasets this considerably reduced their sizes. Henselman et al. [13] concludes that unlike the PER there are no clear trends for the WER. It seems that the phoneme language models fared better in comparison to the word language model. The results can most probably be improved with a larger corpus since the Lwazi is smaller than the NCHLT corpus. An example of a language model that was developed and applied, created subtitles for Japanese broadcast news [1]. The language model was developed using a database of news manuscripts created by NHK, Japan's broadcasting corporation. The news manuscripts were divided into morpheme units, which is when a word is broken up into the smallest units that make up the word. These morpheme units were then used as the training data for the language model since Japanese sentences are not divided word for word like in English. Words were then selected according to how many times they appeared and then registered in the pronunciation dictionary. A further adaptation of this language model was created whereby a final reviewed and edited news manuscript could be used for training data since little to no errors would occur in this final manuscript. Hence, this would improve speech recognition performance.

6. Pronunciation dictionary

A pronunciation dictionary/lexicon can be described as a dictionary of words and their pronunciations. Grapheme-to-phoneme (G2P) is a process of using rules to create a pronunciation dictionary [20]. Besacier et al. [5] proposed some options on how to create a pronunciation dictionary for under-resourced languages. The options are, to use a knowledge-based approach, an automatic approach, or a grapheme-based approach. The knowledge-based approach is used to design a phonetizer which is time-consuming and requires satisfactory knowledge of the language. The grapheme-based approach describes the words in terms of grapheme units. Nevertheless, in the context of under-resourced languages, there are many challenges [4]. It is argued that other phonetic sounds such as affricates, diphthongs and click sounds, commonly found in South African languages, could be modelled as single units or sequences, and allophones. A study done by van Niekerk et al. [20] used a corpus divided into three categories: regular pronunciations, irregular pronunciations, and pronunciation addenda, which are speaker-specific pronunciations. The isiXhosa, Sesotho and Setswana dictionaries were created automatically by using the G2P rules from the NCHLT corpus. Since these languages are more phonemic in their orthography, less manual correction was required, unlike Afrikaans. In contrast, there were inaccuracies caused by the South African and Lesotho orthographies of Sesotho which had to be manually corrected. For pronunciation verification, flagged words needed to be manually reviewed. Since these tasks had to be done manually, quickly, and effectively, only particular categories were reviewed. Examples of pronunciation dictionaries include, Tarján et al. [19] who used an automatic approach and created a pronunciation dictionary for Hungarian words that were pronounced differently from the usual way, Ando et al. [1], with regards to the Japanese language, finalised the pronunciation dictionary through human confirmation, unlike Tarján et al. [19], since, depending on the context, the pronunciation of the word in Japanese may be different. Although a disadvantage to creating a pronunciation dictionary is that it potentially requires manual verification, these dictionaries help create an accurate ASR system.

7. Summary

The nine low-resource South African Bantu languages have very few electronic textual resources. Creating these resources poses several challenges. Fortunately, a substantial amount of research has been done to aid in the transcription to create these resources. It was seen by comparing various toolkits that the Kaldi toolkit performed the best for the South African languages. Although, a few evaluations of low-resource languages said hybrid DNN systems exhibited preferable results, these models perform poorly with minimal training data. The research is done on 155 minutes of the language Seneca using a DNN the results were unusable. Whereas when the Kaldi toolkit was used, yielded much better results. Since the NCHLT has more than 50 hours of audio it possibly can be used in a DNN system, although this would need to be tested. Further research would have to be done on DNN for the transcription of South African languages since it is unclear what amount of training data is required for a DNN system to achieve positive results. Thus, from the research, it can be deduced that for the South African languages, Kaldi would be the better toolkit to use. Furthermore, numerous methods are used for training the data. The machine learning techniques that are favoured are unsupervised and semi-supervised training. As the South African languages lack textual resources for training data, these methods are recommended. Audio partitioning is further used to divide acoustic signals into segments. This helps reduce computation time. Many advantages are highlighted by using segmentation such as identifying speaker turns and identities as well as identifying and removing non-speech segments. Segmentation is therefore exceptionally useful in broadcast news where a change in speaker turns happens often. This is also known as speech diarization. Many papers that were reviewed on the transcription of broadcast news used speaker diarization since it was useful to partition the speech into segments which correspond to speaker turns. Observing the use of speaker-independent and speaker-dependent segmentation showed a higher accuracy in speaker-dependent segmentation with a lower WER and PER. Therefore, a faster transcription speed could be reached. An increased accuracy could also be achieved by data augmentation. This was verified by augmenting data 25 times which yielded a significant decrease in WER. Although this seemed to work for a DNN model when using the Kaldi toolkit, the original unaugmented data had a lower WER than the augmented data. In terms of the South African languages, language models were compared and evaluated using perplexity. The perplexities of the nine ethnic languages were lower than that of Afrikaans suggesting that these languages have a more predictable phoneme sequence. Among comparing the phoneme and word language models it was seen that the phoneme language model outperformed the word language model. An example showed that a language model was created to successfully transcribe Japanese broadcast news by using the news manuscripts as training data. Lastly, options on how to create pronunciation dictionaries were proposed where it was seen that an automatic approach worked in creating a dictionary for isiXhosa, Sesotho, and Setswana. These languages also required less manual correction to their dictionaries than Afrikaans. Transcribing SABC broadcast news in the nine ethnic South African languages is possible considering the previous research done. There are open-source toolkits available as well as different methods in improving accuracy and performance such as segmentation and augmentation to create an accurate ASR system for the transcription of SABC news.

REFERENCES

[1] A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, “Real-time transcription system for simultaneous subtitling of Japanese broadcast news programs,” *IEEE transactions on broadcasting*, vol. 46, no. 3, pp. 189-196, 2000.

[2] E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraçlar, “Turkish broadcast news transcription and retrieval,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 874-883, 2009.

[3] E. Barnard, M. Davel, C. van Heerden, F. De Wet, and J. Badenhorst,"The NCHLT speech corpus of the South African languages," *4th International Workshop on Spoken Language Technologies for Under-Resourced Languages,* pp. 194-200, 2014.

[4] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey,” *Speech communication*, vol. 56, pp. 85-100, 2014.

[5] L. Besacier, V.-B. Le, C. Boitet, and V. Berment, “ASR and translation for under-resourced languages,” *IEEE*, 2006.

[6] F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani, “A baseline for the transcription of Italian broadcast news,” IEEE, pp. 1667-1670, 2000.

[7] G. D. Cook, D. J. Kershaw, J. Christie, C. W. Seymour and S. R. Waterhouse, “Transcription of broadcast television and radio news: The 1996 ABBOT system,” IEEE, 1997.

[8] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, “The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news,” 2005.

[9] R. Eiselen, and M. J. Puttkammer, “Developing Text Resources for Ten South African Languages,” pp. 3698-3702, 2014.

[10] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, “Speech recognition and keyword spotting for low-resource languages: Babel project research at cued,” *International Speech Communication Association (ISCA)*, 2014.

[11] J.-L Gauvain, L. Lamel, and G. Adda, “Transcribing broadcast news for audio and video indexing,” *Communications of the ACM*, vol. 43, no. 2, pp. 64-70, 2000.

[12] V. Gupta, and G. Boulianne, “Speech transcription challenges for resource constrained indigenous language Cree,” *Proceedings of the 1st Joint SLTU and CCURL Workshop*, pp. 362-367, 2020.

[13] D. Henselmans, T. Niesler, and D. Van Leeuwen, “Baseline speech recognition of South African languages using Lwazi and AST,” *in Proceedings of the 24th Annual Symposium of the Pattern Recognition Association of South Africa,* 2013.

[14] R. Jimerson, K. Simha, R. Ptucha, and E. Prud'hommeaux, “Improving ASR output for endangered language documentation,” 2018.

[15] L. Lamel, J.-L. Gauvain, G. Adda, M. Adda-Decker, L. Canseco, L. Chen, O. Galibert, A. Messaoudi, and H. Schwenk, “Speech transcription in multiple languages,” *IEEE*, pp. 757-760 2004.

[16] Naidoo, A. and Tsoeu, M. “Evaluating Open-source Toolkits for Automatic Speech Recognition of South African Languages,” *IEEE SAUPEC/RobMech/PRASA Conference*, 2019.

[17] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, “Speech recognition using deep neural networks: A systematic review,” *IEEE*, vol. 7, pp. 19143-19165, 2019.

[18] T. Patel, D. Krishna, N. Fathima, N. Shah, C. Mahima, D. Kumar, and A. Iyengar, “An Automatic Speech Transcription System for Manipuri Language,” *Interspeech*, pp. 2388-2389, 2018.

[19] B. Tarján, P. Mihajlik, A. Balog, and T. Fegyó, “Evaluation of lexical models for Hungarian Broadcast speech transcription and spoken term detection,” *IEEE*, 2011.

[20] D. van Niekerk, C. van Heerden, M. Davel, N. Kleynhans, O. Kjartansson, M. Jansche, and L. Ha, “Rapid development of TTS corpora for four South African languages,” *Interspeech*, 2017.

[21] N. T. Vu, F. Kraus, and T. Schultz, “Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training,” *Interspeech,* 2011.

[22] J. Zhao, and W.-Q. Zhang, “Improving Automatic Speech Recognition Performance for Low-Resource Languages with Self-Supervised Models,” *IEEE Journal of Selected Topics in Signal Processing*, vol.16, no. 6, pp.1227-1241, 2022.