The semi-automated creation of stratified speech corpora

Charl van Heerden, Marelie H. Davel and Etienne Barnard
Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa
cvheerden@gmail.com

***Abstract*—Smartphones provide an efficient means for the collection of speech data; however, the quality of the corpora created in this fashion is not predictable. We describe an approach that allows us to post-process and rank utterances in a prompted speech corpus quickly and effectively. Utterance ranking makes it possible to both select those utterances with the highest likelihood of being correct and to evaluate the quality of the resulting corpus from a limited sample. This approach has been applied to a collection in the eleven official languages of South Africa, and we show that it naturally leads to the creation of stratified corpora from the same collection. Such corpora can be useful for different purposes, and corpus users are provided with the tools to extract these easily: from small, highly accurate corpora to larger corpora that are likely to contain more errors.**

Index Terms: speech corpora, automatic speech recognition, confidence scoring

I. Introduction

The widespread availability of internet connectivity and smartphones has stimulated the development of novel approaches to data collection [1], [2], which use the smartphone device both to prompt participants with text and to record the resulting speech. This approach can lead to the rapid and efficient collection of substantial amounts of speech, along with the prompting text that can be viewed as baseline transcriptions for these recorded utterances.

Experience with smartphone data collection in several languages has shown that speech corpora collected in this fashion are directly usable for the training of automatic speech recognition (ASR) systems [1], [2]. Even though various factors may cause deviations between the prompted text and the recorded speech, such differences are generally small enough not to disrupt the training of acoustic models for ASR [2]. However, there are several reasons why it may be preferable to create a corpus with a more controlled level of audio and transcription quality:

- In order to evaluate the performance of ASR and other systems, the transcriptions serve as "ground truth", and therefore need to be suitably reliable.
- For linguistic studies, for example on the acoustic phonetics of the target language, reliable transcriptions can eliminate confusion and improve the robustness of analyses.
- As the proportion of errors increases (e.g. because of limitations in the literacy levels of participants), we are likely to reach a stage where ASR accuracy can be improved by eliminating or improving inaccurate transcriptions.

Of course, the conventional solution to this problem is to employ human transcription in order to create an orthographically transcribed corpus with very high reliability. However, the cost of such transcription is prohibitive for many resource-constrained applications, and the resulting accuracy may be excessive for the types of applications mentioned above. It is therefore clear that a method to create quality-controlled speech corpora from smartphone-collected data with limited human effort would be of great practical value, especially for application in under-resourced languages.

Below, we first summarize some of the previous research related to this task. Section III describes the input data that was utilized as well as the approach that we have taken. The results obtained with eleven South African languages are summarized in Section IV. In the concluding section we discuss some of the implications of this work, and suggest extensions and improvements.

II. Background

As motivated above, the prompts that are provided to participants during data collection cannot be used directly as transcriptions for a high-reliability corpus, since discrepancies between recorded audio and the corresponding prompt text arise from factors such as reader errors or disfluencies, background noise, and issues with the recording device. Hence, our goal was to convert this list of recordings and associated prompts into a usable corpus by two main mechanisms:

- The elimination of those recordings that clearly do not correspond to the prompted text.
- The addition of noise markers to the transcriptions of recordings that contain additional sounds beyond the prompted text.

The first task is essentially a case of *confidence scoring*, which has been studied in great detail in the speech-recognition literature. Several of these approaches are summarized in [3], where it is also shown that this particular instance of confidence scoring has somewhat unique characteristics compared to other applications that have appeared in the literature. In [3] an algorithm that we call phone-based dynamic programming (PDP) is shown to be particularly well suited to the current task.

PDP assumes that the corresponding prompt-utterance pairs contain exactly the same phrases, and scores how accurate the match between audio and text is. However, we often find that the recordings differ from the prompts in predictable ways - for example, the speaker may stutter, or repeat a word, or background noise may intrude into the recording. Such utterances will, correctly, be rejected by PDP in general, since the discrepancy between text and audio is substantial. However, these utterances contain much useful speech, and we realized that we could recover that speech (for most speech-processing purposes) by marking the extraneous sounds (which could arise from speech or some other acoustic source) as "noise". We have developed a specialized *garbage model* for this purpose, which is briefly summarized in Section III and motivated in more detail in [4].

III. INPUT DATA AND APPROACH

Our development started with *Woefzela*-collected speech in the official languages of South Africa, which had been developed at the Meraka Institute[2]. Speakers were prompted to read approximately 500 short prompts each, which had been extracted from text corpora in order to maximize variability of phonemic contexts present in the text[2]. After some initial pre-filtering, we had a set of baseline corpora, nine of which contained speech from exactly 210 first-language speakers each, whereas the other two languages had somewhat fewer speakers. The statistics of these initial corpora – referred to as the *NCHLT-baseline* corpora from here onwards – are provided in Table I. From these baseline corpora, a set of *NCHLT-clean* sub-corpora are selected: the selection process aims to select utterances that are likely to be transcribed with high accuracy ($> 95\%$ word accuracy). (By prior agreement with the sponsors of the corpus, we aimed for clean corpora of 50 to 60 hours of speech per language.)

TABLE I. SUMMARY OF BASELINE CORPORA (*NCHLT-baseline*) USED FOR DEVELOPMENT; CORPUS DURATIONS ARE IN HOURS.

Language	Speakers	Males	Females	Duration
Afrikaans	210	107	103	100.6
English	210	100	110	87.0
isiNdebele	148	78	70	101.8
isiXhosa	210	107	103	165.0
isiZulu	210	98	112	157.2
Sepedi	210	100	110	122.6
Sesotho	210	113	97	133.5
Setswana	210	109	101	128.3
Siswati	198	96	102	139.3
Tshivenda	210	84	126	154.8
Xitsonga	200	95	105	142.6

As discussed above, we employed *garbage modeling* and *PDP scoring* in order to insert "noise" markers and select well-spoken utterances (relative to the prompted text), respectively. Both of these approaches require a speech-recognition system for their operation, and we employed the HTK toolkit [5] to develop baseline recognizers in each of the languages, training on a randomly selected subset of 190 speakers (but 128 for isiNdebele and 178 for Siswati) and using the other 20 speakers as development set for parameter tuning and performance monitoring. A standard 3-state left to right HMM architecture was used to model context-dependent triphones in each language. As acoustic features, 39-dimensional Mel Frequency Cepstral Coefficients were used: 13 static coefficients with cepstral mean normalization applied, 13 delta and 13 double delta coefficients. Triphones were tied at the state level using decision tree clustering, and each tied-state triphone was estimated with 8 Gaussian mixtures per state. Semi-tied transforms were also employed throughout.

The *garbage model* used during alignment is based on a background model that can be inserted between any pair of words. The garbage model is a 3-state global HMM, with 16 mixtures per state. Apart from the number of mixtures, it is trained using the same parameters and features as models of the general recognizer, but on all the data (that is, an independent training cycle, using the same data as the general recognizer). After initial training, this model is then extended by adding a short pause model in parallel. This model is implemented as an HTK tee-model (free transition from entrance to exit state), with transitions allowed to, from and between the 3-state global model and the fourth short pause state. The result is a general model which can absorb large spoken sections and/or silence, or can be skipped completely.

During corpus development, we employ our initial acoustic models (in each language) to perform forced alignment between the prompted text and the recorded utterance. (The pronunciation dictionaries used for this purpose were extended versions of those described in [6] and related publications.) Frames that are matched by the garbage model are marked as such, and if the "garbage" frames correspond to at least five phones in a standard phone-loop decode (during subsequent PDP scoring), a noise symbol is inserted into the transcription. Not all noise markers initially marked are therefore retained: subsequent PDP scoring is used to flag valid instances of noise and/or partial words, as described in more detail later in this section.

These forced alignments, with optional noise markers, are also used by the PDP scoring algorithm. For this step, we perform phone recognition with an ergodic phone loop of the same audio segment, and compare the phone strings with the forced alignments to obtain a distance measure. We use dynamic programming – with either a flat or a variable scoring matrix – to map the one phone string to the other. The variable scoring matrix allows us to penalize more probable recognition errors less severely than differences that are more likely to be indicative of an unmatched text/audio segment. We then normalize the resulting dynamic programming score and use this as confidence score. While this measure is less grounded in the standard Bayesian theory of ASR than those that estimate a posterior probability of the presumed transcription, it is likely to be less fragile in environments where the acoustic models do not match the acoustics of the target utterances very well [7]. In detail, then, the PDP scoring process consists of the following steps:

1) Free recognition is performed on the audio segment using a phone-loop grammar in order to produce an *observed string*.
2) An ASR-based alignment of the prompted text, with possible inserted noise markers, produces a *reference string*.
3) The phone set is simplified, and both the observed and reference string are mapped to the simplified phone set. Specifically, compound phones (such as affricates or diphthongs) are split into their constituent parts, and a few rare phones mapped to their closest counterparts.
4) A standard dynamic programming algorithm (with a pre-calculated scoring matrix - see [3] for details) is used to align the observed and reference string with

each other. Noise symbols can be matched against any phoneme at zero cost, since those may correspond to arbitrary insertions.

5) The resulting score obtained from the best dynamic programming path is normalized by the number of phones in the alignment (ignoring noise symbols).

Related algorithms have been proposed by several authors – possibly the earliest systematic exploration of this class of algorithms was performed by Ng and Zue [8].

This process – training of initial acoustic models, detection of potential discrepancies between text and audio with a garbage model, and scoring of utterances with PDP – was executed on each of the eleven baseline collections. Of course, this process could be iterated: the cleaned corpus could be used to train new acoustic models, which could be used to repeat the subsequent segmentation and scoring steps. However, we have found that changes after the first iteration are relatively small if reasonable initial transcriptions are employed [4]; hence, it is doubtful whether additional iterations would be worthwhile.

We next summarize various measurements that describe the corpora that were developed using these methods.

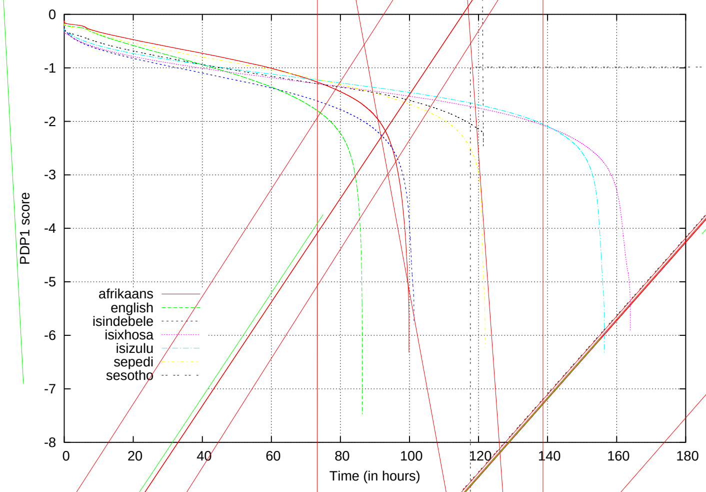
IV. Measurements of corpus size and quality

The PDP scoring algorithm is designed to provide a rank ordering of recorded utterances: utterances with larger scores are more likely to contain dictionary-expected enunciations of the prompted text, and as scores become progressively lower, errors are increasingly likely to occur. We have verified this to be the case (see [3] as well as Figs. 2 and 3 below). Hence, the duration of utterances that exceed a given PDP score is a good indication of the amount of speech that meets or exceeds the corresponding quality measure, where “quality” accounts for both the acoustic quality of the utterance and the match between the speech and the transcribed text. Fig. 1 shows this curve for each of the languages in our corpora. All these graphs show comparable behaviors: the number of utterances above a certain threshold increases gradually as the threshold is decreased, and in each language there is a much smaller “tail” of utterances with very low PDP scores (-2.5 or less). These utterances should definitely not be included in the corpus, as they are likely to contain one or more of the errors described in Section III. We also see that the targeted corpus size (50 to 60 hours of recordings) places us well outside of the low-quality tail in each of the languages. We decided to retain approximately 56 hours of speech in each language; that allows us to accept only utterances with PDP scores above approximately -1.3 in English and isiNdebele, with larger PDP thresholds being employed in the other languages.

With this selection process complete, it remains to determine how accurately the selected utterances were produced. As we discuss below, the definition of “accuracy” in this context is rather complex; we have therefore developed an evaluation protocol that allows us to arrive at reasonable estimates, using a set of carefully annotated evaluation utterances.

A. Evaluation protocol

For the evaluation process, we manually validated 400 randomly selected utterances in a sampling of languages. As



additional words, and no poorly pronounced words or reading errors. In this case, the only difference between "strict" and "harvest" scoring is that utterances containing close matches are rejected or accepted, respectively. Insertions are deemed acceptable if the garbage model is employed.

Corpus validation is performed by accepting all words that exceed a selected threshold and determining to what extent this automatic accept/reject decision matches the manually generated accept/reject decision. As we can trade off between sensitivity and specificity by adjusting the automatic accept/reject threshold, we evaluate the effectiveness of our scoring technique for each threshold setting, and first use a Detection-Error Trade-off (DET) curve to plot the fraction of correctly detected acceptable words against the percentage of correctly rejected unacceptable words. (The closer to the top right-hand corner the curve, the more effective the technique.)

B. Results

Figs. 2 and 3 demonstrate the performance achieved for isiNdebele and Afrikaans. As the PDP threshold is increased, corpus selection becomes stricter, identifying and rejecting an increasing number of true errors. At the same time, less and less of the correct portion of the corpus is accepted. Curves are shown for different types of PDP scoring: using a flat matrix or a trained matrix, and employing strict or harvest scoring. (All curves are shown for scores extracted after phone splitting and before the garbage model is employed - these generate fairly similar DET curves.) We see that there is some variability between the acceptance / rejection performance achieved in the different languages: for example, when using strict scoring in Afrikaans we are able to reject almost 90% of the erroneous utterances, while still retaining about 70% of the correct utterances, whereas a similar rejection level in isiNdebele entails that somewhat more than 40% of the correct utterances will be retained. Similarly, when using harvest scoring, 90% of erroneous utterances are rejected while retaining almost 90% of the correct utterances in Afrikaans, and 75% of the correct utterances in isiNdebele.

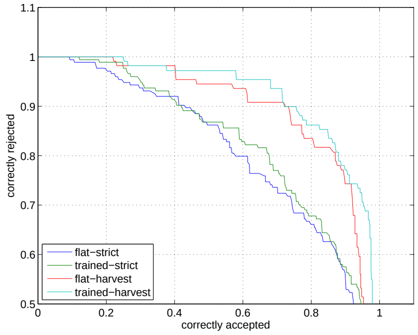


Fig. 2. DET curve for isiNdebele using strict or harvest scoring, and employing either a flat or trained PDP matrix.

Fortunately, the fraction of correctly-recorded utterances is sufficiently high for good word accuracies to result from the levels of rejection performance that we have achieved. At the thresholds used for corpus selection, very high accuracies are observed: approximately $97.3\% - 98.5\%$ when using strict scoring, $99.4\% - 99.7\%$ when using harvest scoring. Table II lists the word accuracies that were achieved for the above-threshold utterances in our various evaluation sets.

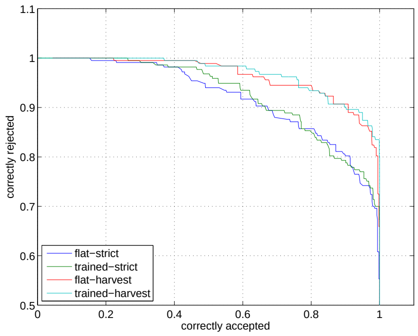


Fig. 3. DET curve for Afrikaans using strict or harvest scoring, and employing either a flat or trained PDP matrix.

TABLE II. WORD ACCURACIES OF SAMPLED UTTERANCES, ACCORDING TO TWO DIFFERENT DEFINITIONS OF "CORRECTNESS"

Language	Strict	Harvest
Afrikaans	97.80	99.38
English	97.68	99.40
isiNdebele	97.30	99.56
Sepedi	98.34	99.60
Sesotho	97.76	99.47
Setswana	98.54	99.74

Based on the analysis described above, the following corpora were prepared for public release by the South African Resource Management Agency:

1) *NCHLT-clean*: the eleven 56-hour corpora described in Section III.
2) *NCHLT-baseline*: the full baseline corpora, also described in Section III, with corpus statistics as in Table I.
3) *NCHLT-raw*: the total set of usable data collected, including repeated speakers and utterances.

V. Conclusion

We have shown that a suitable approach to confidence scoring, combined with garbage modeling, can be used to create corpora with well-defined transcription accuracies out of smartphone-collected speech data. Since this approach does not require sophisticated language modeling, it is particularly suitable for the development of speech corpora in under-resourced languages. Using a strict definition of accuracy, we estimate that the word error rates in our transcriptions range between 1.5% and 2.7%. Clearly, these values depend on the details of the collection process, prompting materials, participant population, and other factors. Thus, the accuracies that will be achieved if similar methods are employed in other circumstances will be somewhat unpredictable, and it will generally be necessary to label a small set of utterances manually in order to estimate the accuracy achieved. However, this "gold standard" set can be much smaller than the overall corpus – we have employed sets of 400 utterances per language, and these could each be scored in less than an hour.

The approach we have described is unbalanced between insertions and deletions of speech: with the garbage model, we are able to detect and correct for inserted speech, but deleted segments of the prompted text are simply marked as errors. This is reasonable in our collection, since insertion errors are more frequent, but it should be reasonably straightforward to increase the efficiency of harvesting without compromising quality by detecting deleted words or even partial words.

It would be interesting to investigate the characteristics of our approach under significantly different circumstances. In particular, we have not seen any benefit to excluding any utterances from ASR training, but it is clear that such benefit will be available if the fraction of errorful recordings becomes large. With data collections in more challenging environments, it should be possible to study this transition. We also look forward to seeing how our corpora are used by speech technologists and linguists; such usage will hopefully give us further guidance on how to adjust our methodology.

References

[1] T. Hughes, K. Nakajima, L. Ha, P. Moreno, and M. LeBeau, “Building transcribed speech corpora quickly and cheaply for many languages,” in *Proc. Interspeech*, Makuhari, Japan, September 2010, pp. 1914–1917.

[2] N. De Vries, J. Badenhorst, M. Davel, E. Barnard, and A. De Waal, “Woefzela - an open-source platform for ASR data collection in the developing world,” in *Proc. Interspeech*, 2011.

[3] M. Davel, C. van Heerden, and E. Barnard, “Validating Smartphone-Collected Speech Corpora,” in *Spoken Languages Technologies for Under-Resourced Languages*, Cape Town, South Africa, May 2012, pp. 68–75.

[4] M. Davel, C. van Heerden, N. Kleynhans, and E. Barnard, “Efficient Harvesting of Internet Audio for Resource-Scarce ASR,” in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 3153–3156.

[5] S. Young, G. Evermann, M. Gaels, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, “The HTK book (for HTK version 3.4,” March 2009.

[6] M. Davel and O. Martirosian, “Pronunciation dictionary development in resource-scarce environments,” in *Interspeech 2009*, 2009, pp. 2851–2854.

[7] E. Barnard, M. Davel, C. van Heerden, N. Kleynhans, and K. Bali, “Phone recognition for spoken web search,” in *Proceedings of the MediaEval 2011 Workshop*, 2011.

[8] K. Ng and V. W. Zue, “Subword-based approaches for spoken document retrieval,” *Speech Communication*, vol. 32, no. 3, pp. 157–186, 2000.