Efficient harvesting of Internet audio for resource-scarce ASR

Marelie H. Davel[2], *Charl van Heerden*[12], *Neil Kleynhans*[12] *and Etienne Barnard*[2]

[1]Human Language Technologies Research Group, CSIR Meraka Institute, Pretoria, South Africa
[2]Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa
{marelie.davel,cvheerden,ntkleynhans,etienne.barnard}@gmail.com

Abstract

Spoken recordings that have been transcribed for human reading (e.g. as captions for audiovisual material, or to provide alternative modes of access to recordings) are widely available in many languages. Such recordings and transcriptions have proven to be a valuable source of ASR data in well-resourced languages, but have not been exploited to a significant extent in under-resourced languages or dialects. Techniques used to harvest such data typically assume the availability of a fairly accurate ASR system, which is generally not available when working with resource-scarce languages. In this work, we define a process whereby an ASR corpus is bootstrapped using unmatched ASR models in conjunction with speech and approximate transcriptions sourced from the Internet. We introduce a new segmentation technique based on the use of a phone-internal garbage model, and demonstrate how this technique (combined with limited filtering) can be used to develop a large, high-quality corpus in an under-resourced dialect with minimal effort.

Index Terms: speech recognition, under-resourced languages, garbage modeling

1. Introduction

The limited availability of speech corpora is a major constraint on the development of automatic speech recognition (ASR) in under-resourced languages and dialects [1, 2]. Consequently, there is significant interest in ways to develop such corpora efficiently [3], and the efficient exploitation of limited corpora [1, 4]. We investigate an alternative source of ASR data for resource-scarce languages, namely speech transcribed orthographically for purposes other than ASR development. Such data are quite common in many languages, since transcriptions are often compiled for purposes such as captioning (open or closed), presentation in printed format, Internet searchability, etc.

The major challenge with transcriptions that were not prepared for ASR purposes is that they are usually a somewhat loose reflection of what was actually said. Disfluencies, repetitions, grammatical errors and the like are generally not transcribed, and errors (especially in technical terms and proper names) and inconsistent transcription conventions are common. Large sections of text may not be transcribed at all, and similarly, additional text not found in the audio data may be included in the transcriptions. We refer to these as "approximate transcriptions", even though they may be perfectly good for human consumption. The benefit of using such corpora is that they are typically quite large by ASR standards, since it is relatively inexpensive to record and provide approximate transcriptions of many hours of speech; and are often freely available, for example, being posted on public Web sites.

For well-resourced languages, the task of corpus development from approximately transcribed speech has received significant attention, especially in the broadcast-news and lecture-transcription domains [5, 6]. Much of this research builds on the availability of a well-trained ASR system and a suitable language model in the language of interest, an assumption which is generally invalid for resource-scarce languages. The main goal of our research is therefore to see whether it is possible to harvest audio corpora from approximately-transcribed speech in the absence of a directly relevant ASR system. For practical reasons, it is also important that the approach developed should not require extensive manual intervention, since the expertise for such interventions can be lacking in under-resourced languages.

The particular task we focus on is the development of a broadband recognizer in South African English (SAE). SAE is a resource-scarce dialect, which is sufficiently distinct from major dialects (such as those of the USA or UK) to present substantial recognition challenges [4]. We base our research on a corpus of recordings that are made available along with approximate transcriptions on a public Web site. We investigate a number of approaches requiring differing amounts of manual intervention and insight into ASR development, and find that it is possible to achieve good results with minimal manual processing and limited specialized (task-specific) ASR processing.

2. Background

The use of minimal or approximate transcriptions for ASR purposes, also referred to as "lightly supervised acoustic training" [7], typically consists of three general steps: (1) data segmentation, (2) word-based alignment and (3) filtering.

During *data segmentation*, heterogeneous data sources are separated based on characteristics such as bandwidth, gender and/or speaker [6, 7], and non-speech segments removed to the extent possible [7]. While Moreno *et al* [8] did not perform data segmentation, such an initial phase is typical to most approaches. Once data has been segmented into fairly homogeneous sub-parts, appropriate acoustic models are trained using existing ASR corpora. In experiments reported on in literature, these base corpora ranged from about 24 hours [6] to 150 hours of speech [9]. In [7] an hour of manually corrected transcriptions was used to bootstrap acoustic model training.

In all of the above cases, acoustic models are then used to perform *word-based alignment*: audio is recognized using a language model strongly biased towards the available transcriptions [5, 6, 7, 9]. Additional cycles of acoustic model adaptation based on earlier recognition results may be performed [9], resulting in a final set of ASR hypotheses. When sequences are very long, a more complex multi-stage alignment process can be followed, based on the identification of reliable anchor points [5, 8].

Finally, some form of *data filtering* is used, either during alignment or upon completion to identify problematic segments (for example, unmatched audio or transcriptions). Standard word error rate is a popular confidence measure used for this purpose, with the ASR hypotheses matched against the approximate transcriptions [7, 6, 9]. Hazen goes one step further in [5], and attempts to rectify problematic transcriptions automatically.

While not specifically utilized in the context of alignment of approximate transcriptions, *garbage models* are frequently used in ASR systems to model spontaneous events (both speaker- and background noise-related). These models are either added to absorb non-speech events explicitly marked in a transcription, or are added between any word pair as an optional event [10].

3. Approach

Our approach to the development of an ASR corpus from approximate transcriptions does not require a data segmentation phase, and relies on an acoustic garbage model during alignment and filtering. (No word recognition is performed and no language model required). Below, we describe our process in more detail.

3.1. Harvesting data from the Internet

Since our goal was to develop a wideband recognizer for SAE, we searched the Internet for publicly-available sources of transcribed speech in that dialect, and contacted the owners of the respective Web sites in order to obtain permission for the use of their data. Of the options available, we selected Moneyweb, a provider of South African business, financial and investment news. Moneyweb provides daily radio broadcasts, which are made available, along with approximate transcriptions and some metadata, on their web site (www.moneyweb.co.za) soon afterwards. These broadcasts contain news reports, interviews and in-depth discussions of current topics. Besides the two main presenters, four journalists make regular contributions, and a large number of guest speakers (some in studio, others calling in) also appear in the broadcasts. These speakers speak in a range of English dialects: the majority of the speech is in standard SAE, but substantial portions are also in other dialect variants (e.g. Zulu-accented or Afrikaans-accented SAE). We downloaded recordings (in mp3 format), along with their approximate transcriptions, corresponding to about two years of broadcasts.

There are a number of challenges to using this data for ASR development: (1) A wide variety of content, including different dialects and speaking styles, but also music and other non-speech content, occurs in the recordings. The majority of the speech is spontaneous in nature, ranging from well-articulated spontaneous speech by professional journalists to hesitant speech in broken English by some interviewees. (2) Since much of the recorded material consists of dialogs, phenomena such as speaker overlap, filled pauses and back-channel communication are frequent. (3) Most of the content consists of wideband speech (8 kHz or more), but telephone-bandwidth recordings are also common; sampling rates range between 16 kHz and 44.1 kHz, and the encoding quality is also somewhat variable; and (4) The transcriptions are of typical "professional" quality: disfluencies, repetitions and filled pauses are omitted, some grammatical errors have been corrected, transcription errors occur and transcription inconsistencies (for example in transcribing dates or abbreviations) are common. These phenomena make it clear that the downloaded recordings are not directly suitable as an ASR corpus. Fortunately, the amount of data available is substantial (we easily retrieved more than 100 hours of speech), which makes it feasible to extract a portion of the data that is suitable for use in ASR.

3.2. Bootstrapping a phoneme recognizer

Since we are not aware of the existence of any broadband SAE corpus, we bootstrap an SAE recognizer while developing the corpus, starting from a US English recognizer. For this purpose, the widely-available Wall Street Journal (WSJ) corpus was utilized. As a starting point, a standard tied-state, context-dependent (triphone) Hidden Markov Model (HMM) recognizer was trained using the HTK toolkit [11]. Each HMM consists of 3 states, with a Gaussian Mixture Model (GMM) with 8 mixtures per state used to model the acoustic data. Models are trained on 39-dimensional Mel-frequency cepstral coefficients (13 static, with their deltas and double deltas), with cepstral mean normalization and semi-tied transforms applied.

In order to apply this recognizer to the SAE recordings, a suitable phone set and dictionary were required. We based our work on the dictionary described in [12], which was bootstrapped from a British English dictionary, using simple rewrite rules. To use this dictionary in conjunction with the WSJ-trained recognizer, we developed a WSJ to SAE phone mapping. We simplified both phone sets by merging phonemes where differences are not consistently modeled in the different source dictionaries. Thus, all diphthongs were split into their constituent monothongs, and diacritics related to stress and duration (or the tense/lax distinction) were removed. All other phonemes were manually mapped to their closest candidate. The resulting phone set did not contain three phonemes that exist in SAE but not in US English. Forced alignment with the WSJ-trained model using this adjusted phone set formed the starting point for our corpus-development process.

3.3. Iterative alignment, filtering and training

Given a set of acoustic models and a pronunciation dictionary, our basic approach is to perform forced alignment of the approximate transcriptions to the recorded speech, also inserting "garbage" markers as required in order to allow for inaccuracies in the transcription. A simple metric is then employed to select portions of the speech that are aligned well; these are used to retrain the acoustic models (using either MAP adaptation or Baum-Welch training from scratch) and used as starting point for additional cycles of the same process. During the first Baum-Welch retraining cycle, the phoneme set is expanded to the full SAE set and the initial bootstrapping models (using the reduced phoneme set) are discarded.

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 4^{th} short pause state. The result is a general model which can absorb large spoken sections and/or silence, or can be skipped completely.

Because we were concerned about the relatively large mismatch between the recordings and transcriptions, and realized the inadequacy of our initial ASR system, we also experimented with a pre-segmentation process. During this process, members of our team manually segmented a small portion of the corpus by speaker turns, removing non-speech portions of the recordings and marking salient phenomena (e.g. the presence of narrowband speech). In Section 4.3 we report on the effect of such pre-segmentation on the quality of the corpus developed. Note that the manually segmented data was not used during the main alignment process discussed in section 4.2.

Once the alignments have been obtained, bad segments of the corpus are identified and automatically removed. During the alignment process HMM state information is retained: this is used to differentiate between true silence (allowed to occur) and garbage audio (removed), both of which are absorbed by our garbage model. Audio sections between portions marked as garbage are evaluated, and if very small (less than 500ms) these are discarded. This is a quick and efficient way to filter the data.

3.4. Packaging the corpus

Once the alignment-selection-training cycles have stabilized, we are able to select the portions of speech that align without any difficulty, and use those as the basis for the final ASR corpus. In particular, we carry out the following steps: (1) The most reliable portions are identified; (2) these segments are labeled using specialized classifiers trained to identify different types of data; and (3) the selected portions of speech along with relevant meta-data (such as detailed timing information, speaker identity and channel identity) are added to the corpus.

In order to identify the most reliable portions of the corpus, a more sophisticated confidence scoring algorithm is used than the time-base filtering used during alignment. Specifically, we decode the audio using our final model and a phone-loop grammar, and also obtain a forced alignment of the same audio. On a per segment basis, we then compare the two phone strings using dynamic programming (DP) and a variable cost matrix. This cost matrix is derived from the confusion matrix and penalizes mismatches less severely if phones tend to be confusable, but is still heavily biased towards finding matched alignments. Since the DP score (the cost of aligning the two strings) is a reliable indicator of the extent to which the two strings match, we use this as our confidence measure for filtering data. By adjusting this threshold, more or less aggressive pruning of the corpus can be obtained.

While it is not strictly necessary to tag the various types of data available in the corpus, we want to isolate telephone-bandwidth speech from the rest of the broadband corpus for future use. (While the entire corpus is recorded as broadband speech, there are telephone interviews contained in the recordings.) To do this, we train a GMM classifier on segments of labeled broad- and narrowband speech. This classifier is then used to classify all 25ms frames as either broad- or narrowband. Complete segments are classified based on a majority vote from the constituent frames. Audio segments are also attributed to specific speakers. The speaker tags are obtained from the approximate transcriptions (where speaker changes are indicated), with the time intervals obtained from the forced alignments.

4. Analysis

To compare the value of the various processing steps that were employed, and to assess the performance of the overall process, we now present a number of detailed results.

4.1. Measuring improvement

Our aim is to measure how well the aligned audio matches the acoustic models, in order to determine the quality of both our corpus and our acoustic models. Since we do not have a gold standard (manually corrected transcriptions) for measuring recognition accuracy, we use a number of proxy measures, both during development and at final evaluation. This makes it possible to monitor the automated harvesting process without requiring any manual intervention. Specifically we track:

- *The acoustic likelihood of the data*: We measure the improvement in the average acoustic log likelihood of the sections of the corpus identified as containing speech, calculated on a per-frame basis. (Note that an improvement in acoustic likelihood is only meaningful if the amount of data absorbed by the garbage model stays the same, or decreases.)
- *The amount of data considered to be non-speech*: Since any portion of the data that does not match the acoustic models is absorbed by the garbage model, we track this percentage, both in order to validate changes in the acoustic likelihood of the data (see above), and as a quality measure in own right.
- *Dynamic programming scores*: These scores (see section 3.4) provide a direct indication of how well the decoded data match the approximate transcriptions; and
- *Best alignment phone accuracy*: We use the forced alignments from our best model as a proxy for manually corrected transcriptions, and calculate conventional phone recognition accuracy and correctness.

A 3-hour development set (used for testing during corpus development) and a 6-hour evaluation set (used for final verification) were selected and kept separate from the training data. A statistical analysis of the corpus showed that the durations of the audio data contributed by the various speakers are highly non-uniform, being skewed heavily towards a few presenters and frequent in-studio guests. When selecting subsets of the training data (as in section 4.3), care is taken to restrict the amount of data selected from these dominant speakers. Apart from limiting the contributions from five over-represented speakers, further data selection was randomized.

4.2. Bootstrapping and Alignment

The main corpus development process is very simple: apart from resampling some of the audio to ensure a consistent sampling rate, no further pre-processing is performed. The initial unmatched WSJ models are used to bootstrap the first alignment-selection-retraining cycle, and this cycle repeated until the various measures stabilize. Results for three such cycles ('retrain1' to 'retrain3') are provided in Table 1. We report acoustic log likelihoods (log P), percentage of data considered non-speech (non-speech), dynamic programming scores (DPS), phone accuracy percentage (phn acc) and phone correctness percentage (phn cor) using the evaluation set.

Table 1: *Improvements observed during bootstrapping and alignment, reported on the evaluation set.*

model	log P	non-speech	DPS	phn acc	phn cor
WSJ	-87.019	37.62	0.139	36.45	45.22
retrain 1	-79.109	32.78	0.337	53.85	60.93
retrain 2	-78.436	31.08	0.358	55.67	62.22
retrain 3	-78.264	30.76	0.359	56.40	62.84

All the measures are seen to stabilize quickly, resulting in a converged corpus after only three iterations. The original corpus (training data only) contained 99.85 hours of speech; after alignment with the final model a cleaned and aligned corpus of 68.01

hours of speech was retained. Phone accuracy is within the range expected for a corpus of spontaneous speech, and smaller subsets of the corpus (with correspondingly higher recognition rates) can be selected by increasing the quality control thresholds, as discussed in section 3.4. While the garbage model absorbs portions of unnecessary audio (without matching transcriptions), the DP scores identify portions of superfluous text (without matching audio) as well as poorly recorded or articulated sections of speech.

4.3. Effect of additional manual intervention

As mentioned earlier, we did not initially anticipate that a completely automated process would converge so easily. But how good are the results really: could we have done better by using at least a small portion of manually segmented data to improve our initial models? We experiment with this by MAP adapting our initial models with different amounts of manually segmented data (developed as described in section 3.3) prior to retraining.

We find that the reliable data does assist initially, with the amount of improvement over the original baseline models directly correlated with the amount of reliable data added. However, we also find that this improvement does not provide a benefit beyond the first retraining cycle, with the simpler process (described in section 4.2) able to achieve the same performance after a further retraining cycle. Comparative results on the development set are listed in Table 2: since the various quality measures remain strongly correlated (as in Table 1), we only report on acoustic likelihoods (log P) and DP scores (DPS). We observe the same trends when experimenting with additional filtering of data (using DP scores), and find that the retraining process does not require problematic sequences to be cut during alignment; removing problematic sections at the end (when they can most reliably be identified) suffice.

Table 2: *Comparing results when bootstrapping with different amounts of adaptation data.*

	initial		retrain 1		retrain2	
model	log P	DPS	log P	DPS	log P	DPS
WSJ	-84.62	0.197	-77.23	0.402	-76.70	0.413
$\frac{1}{2}$hr MAP	-81.93	0.263	-77.27	0.403	-	-
1hr MAP	-81.17	0.298	-77.19	0.408	-	-
2hr MAP	-80.12	0.336	-77.14	0.410	-	-
4hr MAP	-79.17	0.365	-77.06	0.413	-76.56	0.416

Table 3: *Comparing results when restricting the size of the data set artificially during alignment.*

	initial		retrain 1		retrain2	
model	log P	DPS	log P	DPS	log P	DPS
10 %	-84.62	0.197	-79.17	0.296	-78.47	0.323
20 %	-84.62	0.197	-78.43	0.340	-77.67	0.378
100 %	-84.62	0.197	-77.23	0.402	-76.70	0.413

4.4. Effect of corpus size

A final question we address relates to the size of the harvested corpus. It was clear from the above results that the garbage-based alignment process is able to recover very efficiently, even if starting from a fairly poor acoustic model. To what extent is this phenomenon reliant on the availability of a corpus of this size? In order to investigate this, we repeat the process described in section 4.2 for significantly smaller corpora (10% and 20% of the full training corpus available) and list the results in Table 3. We see that the same trends are observed, even though convergence is slower. (The log P and DPS measures are only listed for the two extreme points evaluated.)

5. Conclusion

We have shown that a sizable ASR corpus can be created from publicly-available resources using a regular and efficient process. This process requires no matching language model, and only partially matching acoustic models. One of our most significant results is finding that that the corpus-development process as defined here does not benefit from the inclusion of any additional manually labeled data. This holds great promise for the development of corpora in under-resourced languages. In the current work, the seed models used to bootstrap the process were from a mismatched dialect, rather than from a different language: we aim to extend this work to determine the extent of mismatch between seed resources and target language that can be tolerated in this fashion.

6. Acknowledgments

We would like to thank Bryan McAlister and his team, who performed the initial data collection and processing; and the South African Centre for High Performance Computing (CHPC), who made their facilities available for these experiments.

7. References

[1] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld, "Healthline: Speech-based access to health information by low-literate users," in *Proc. IEEE Int. Conf. on ICTD*, Bangalore, India, Dec. 2007, pp. 131–139.

[2] E. Barnard, M. Davel, and C. van Heerden, "ASR corpus design for resource-scarce languages," in *Proc. Interspeech*, Brighton, UK, Sept. 2009, pp. 2847–2850.

[3] T. Hughes, K. Nakajima, L. Ha, P. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *Proc. Interspeech 2010*, Makuhari, Japan, Sept. 2010.

[4] C. van Heerden, E. Barnard, and M. Davel, "Basic speech recognition for spoken dialogues," in *Proc. Interspeech*, Brighton, UK, Sept. 2009, pp. 3003–3006.

[5] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Proc. Interspeech*, Sept. 2006.

[6] M. Meng, S. Wang, J. Liang, P. Ding, and B. Xu, "Full utilization of closed-captions in broadcast news recognition," in *Proc. IS-CSLP*, Kent Ridge, Singapore, Dec. 2006.

[7] L. Lamel, J. luc Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[8] P. J. Moreno, C. Joerg, J.-M. Van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Proc. ICSLP*, 1998, paper 0068.

[9] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. Interspeech*, Makuhari, Japan, Sept. 2010, pp. 2222–2225.

[10] L. ten Bosch and L. Boves, "Survey of spontaneous speech phenomena in a multimodal dialogue system and some implications for ASR," 2004.

[11] 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.

[12] L. Loots, M. Davel, E. Barnard, and T. Niesler, "Comparing manually-developed and data-driven rules for P2P learning," in *Proc. PRASA*, Stellenbosch, South Africa, Nov. 2009, pp. 35–40.