INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Building ASR corpora using Eyra

Jón Guðnason, Matthías Pétursson, Róbert Kjaran, Simon Klüpfel, Anna Björk Nikulásdóttir

Center for Analysis and Design of Intelligent Agents, Reykjavik University, Iceland
jg@ru.is, matthiasp@ru.is, r@kjaran.com, simon.kluepfel@gmail.com, annabn@ru.is

Abstract

Building acoustic databases for speech recognition is very important for under-resourced languages. To build a speech recognition system, a large amount of speech data from a considerable number of participants needs to be collected. Eyra is a toolkit that can be used to gather acoustic data from a large number of participants in a relatively straight forward fashion. Predetermined prompts are downloaded onto a client, typically run on a smartphone, where the participant reads them aloud so that the recording and its corresponding prompt can be uploaded. This paper presents the Eyra toolkit, its quality control routines and annotation mechanism. The quality control relies on a forced-alignment module, which gives feedback to the participant, and an annotation module which allows data collectors to rate the read prompts after they are uploaded to the system. The paper presents an analysis of the performance of the quality control and describes two data collections for Icelandic and Javanese.

Index Terms: ASR corpora building, automatic speech recognition, under-resourced languages, speech quality control

1. Introduction

The biggest hurdle to developing automatic speech recognition (ASR) software for a new language are resources. Open-source software for training and running ASR systems has proliferated and there is an ever increasing availability of open-source software that processes audio and text data for acoustic- and language modelling. The hardware for implementing the training and running ASR has also become much more affordable so the entry barrier to ASR is lower than it ever was. Gathering and obtaining language resources therefore deserves greater attention from the speech community than it already has, as it has become the largest engineering obstacle in developing working systems for new languages.

Broadly speaking, the language resources needed for ASR are a large text corpus for language modelling, a phonetic dictionary and an acoustic database with a large number of voice-text tuples for acoustic modelling. This paper presents a system for creating such an acoustic database. We use an open-source system called Eyra, which is designed to have many participants read predetermined prompts from a web-based application [1]. A description of a prototype version of the Eyra system was presented in [1] but here we address the issue of quality control. Collecting voice samples “out in the wild” always means that some percentage of the data will be unsuitable for ASR training. For example, the recording conditions might be unsuitable (e.g. background noise, thumb on microphone) or the participant might be misreading the prompts. A quality control (QC) system is therefore essential for the collection process. The design of Eyra aims to integrate the quality control with the collection procedure. Data collectors are able to rate recordings on-the-fly and participants will get a feedback on how well they are reading.

This paper presents the new updated Eyra which includes an annotation module and an error feedback system. It also presents qualitative results in using the system for gathering acoustic data for Icelandic and Javanese.

1.1. Relation to other work

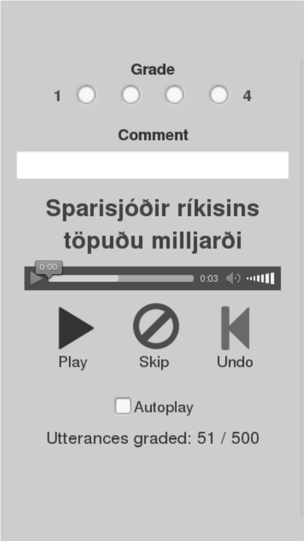
There are numerous examples of speech databases that have been developed for automatic speech recognition. Two well cited databases are the Wall Street Journal Continuous Speech Recognition database [2] and the Librispeech database [3]. Both of these rely on predetermined texts, newspaper articles and books respectively, and participants that read those texts aloud to produce the audio recordings. Other databases are produced by transcribing speech recordings, for example the Switchboard database [4]. These are all examples of databases that have been developed for English but large acoustic databases have been created for many well-resourced languages such as Chinese [5] and Spanish [6]. Developing these databases has required a devotion of resources, both in terms of man-power and expertise. The challenge for under-resourced languages is to emulate this work while minimizing the resources required. Automatic speech recognition for under-resourced languages has received some attention recently. A good survey of the topic can be found in [7].

Prompt-based speech data acquisition has been proposed and developed before [8, 9]. Datahound is an Android based smartphone app that was introduced by Google developers [8]. It has been used to collect acoustic databases for in-house ASR development at Google and it was used to collect an open-source acoustic database for Icelandic called Malromur [10]. The system allows data collectors to download prompt lists to smartphones for recording and subsequent uploading of voice samples. Woefzela was proposed as an open-source alternative to Datahound [9]. It is also an Android-based app and is built to be more robust to sparse data connectivity than Datahound. Woefzela allows field workers to extract the data using external memory cards during data collection sessions. This was deemed necessary for collections where there was no internet connection. Woefzela also included a semi-automatic quality control process that increased the portion of accurately read recordings. Early detection of low volume level and start/stop mistakes makes it possible for the field worker to intervene and correct flaws in the process (e.g. thumb covering microphone).

1.2. Design of Eyra

The design of data acquisition system Eyra is based on the design criteria and experience of Datahound and Woefzela. The convenience of Datahound, the automatic download of prompts and upload of recordings, is maintained in the design of Eyra. A robust back end server contains an SQL database for the prompts, the recordings and all metadata used in the collection, and a front end designed in HTML5 allows for recording

http://dx.doi.org/10.21437/Interspeech.2017-1352



Sparisjóðir ríkisins töpuðu milljarði

Utterances graded: 51 / 500

Figure 1: *The interface of the annotation module.*

through a browser or an app. The design also takes lack of internet connectivity into account by allowing a back end system to be set up locally on a laptop running a local wireless network. The flexibility of the system design therefore allows the data collectors to decide whether the system is run in a crowd-sourcing mode, where data connectivity is not a concern; in a local wireless network mode, where the server is run off a laptop with a fixed number of clients (handsets and/or terminals) for data gathering; or in a hybrid mode where connectivity is not an issue but the number of contributing clients is limited. Details of the server and client design is given in [1]. The rest of the paper describes the design of the annotation module, the error feedback system and data collections for Icelandic and Javanese.

2. Recording and annotation modules

Eyra is designed in such a way that data collection can be organized in many different ways. For example, a crowd-sourcing effort can be set up on a server and the participants can register through a web-based client interface and start the recording. This method means that there is less control over the recording conditions, acoustic environment and the client set-up than if the collection effort is done through a fixed number of client handsets where the data collection team turns up at a location and recruits participants directly.

Quality of the recordings is always going to be of concern regardless of the data collection method used. Main quality concerns during the data collection include muffled or noisy speech (e.g. thumb on microphone or bad acoustic environment), misreading of prompts, and misaligned recording (e.g. if participant presses Next button too soon). The first step in addressing quality in Eyra is the annotation module. This allows data collectors to listen to and rate individual utterances during the collection period.

The annotation mode is reached via a drop-down menu in the client. The annotator chooses a set of utterances to be rated and starts listening. See Fig 1 for a screenshot of the annotation interface. The annotator can choose a grade between one (very poor) and four (very good) and a comment is chosen from a drop-down menu if a grade of one or two is selected. Next prompt appears immediately and the utterance is played automatically if Autoplay is selected. The ratings and the comments are then stored in the relational database alongside the prompts and the recordings.

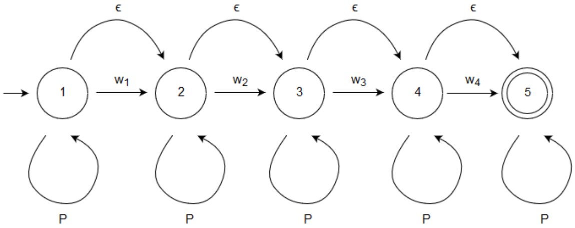


Figure 2: *G FST where P is a phoneme bigram model and the w_i's are the word sequence.*

3. Forced alignment scoring

The aim of the forced alignment scoring is to provide an assessment of how well the spoken utterance matches the corresponding prompt. A simple acoustic model is needed to do this. This can pose a problem since it is quite possible that the data needed to train such a model might not exist. If it doesn't, it is possible to use an acoustic model based on other languages [11] or an acoustic model can be trained on a dataset that is recorded in the start phase of the collection effort. The latter method was chosen for this project. The data collected during this phase does not, therefore, have the error feedback mechanism but the data collected was only about 3 hours in duration. The acoustic model trained for the aligner was developed using the Kaldi speech recognition toolkit. It was based on a simple setup with monophone models using hidden Markov models and 4000 Gaussian mixtures. Training a more complex acoustic model would require more data which is naturally unavailable.

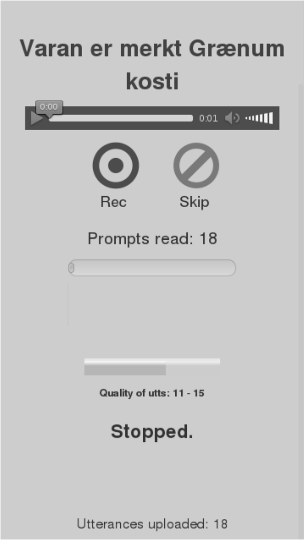
The decoding graph used for the alignment was based on the second alignment stage in the Librispeech corpus building [3] and is depicted in Fig. 2. The graph is formed from the sequence of words in the prompt being aligned and a generic phone bi-gram. The options at each state are to choose the next word, skip the next word or to choose the generic bi-gram and thus inserting an arbitrary amount of phones in the hypothesis. The decoded utterance can therefore have deletions, due to the skipped words, insertions, due to the phone bi-gram and substitutions, due to a combination of the phone bi-gram and a skipped word.

The decoded utterance can now be compared with the prompt. We experimented with phone-, word- and a hybrid phone/word error rate and found that phone error rate was most suitable for this task. The prompts used are typically between one and seven words so the resolution of the word-error-rate (WER) is too low to give a workable error measure. For example the following reference prompt (labelled ref) has been decoded in the following hypothesis (labelled hyp):

```
Jarðskjálftar við Kistufell           ref
jarðskjálftar við !c !I !s !t !Y !t   hyp
```

using word-error-rate, this would count as one substitution and the Levenshtein edit distance therefore gives a $1/3$ WER. Comparing the resulting phonetic transcriptions however:

```
jarðscaultar vIðcIstYf ϵtl            ref
jarðscaultar vIð cIstYt               hyp
```



Varan er merkt Grænum kosti

0:00 0:01

Prompts read: 18

Quality of utts: 11 - 15

Stopped.

Utterances uploaded: 18

Figure 3: *An example of the error feedback provided by the quality control module. The colour of the meter towards the bottom indicates the average score of previous 5 recordings.*

gives one substitution and three deletions and the Levenshtein edit distance of $3/23$ or phone-error-rate (PER).

In addition to giving a higher resolution for scoring, the phone-error-rate also reflects the objectives of the quality assessment better than the word-error-rate. Since the goal is to build a database for ASR acoustic modelling, the aim should be to reward utterances that are phonetically close to the transcript and penalize utterances that don't. As can be seen in the above example, even though the decoding didn't prefer the word `Kistufell`, the phone string it chooses in its place: `cIstYt`, is still close to the transcription of the correct word `cIstYfɛtl`.

4. Error feedback system

Eyra uses phone-error-rate to give feedback to the participant during recording. The overall aim is to increase the quality of the recordings but it is not straightforward how the phone-error-rate is used because the other factor in the equation is quantity of recordings collected. An oversensitive feedback system can easily slow down a recording session and frustrate the participant. This then increases the overall time of the collection effort and/or decreases the collected data quantity. These considerations are taken into account when designing the error feedback system.

The phone-error-rate is converted to a score between zero and by capping the rate at one: Score $= 1 - \min(\text{PER}, 1)$. The forced alignment and this scoring is all done on the back-end server. An average is computed for every five utterances and sent back to the client.

The client quantizes the scores into three categories: 0-0.2 for bad, 0.2-0.7 for medium and 0.7-1 for good. The categories are then presented colour-coded beneath the progress bar on the user interface. Figure 3 shows the user interface during a recording session. The participant has just read the prompt “Varan er merkt Grænum kosti”, she has read 18 prompts and the medium quality of utterances 11-15 is signalled with a yellow meter towards the bottom of the screen. This meter will turn green for good utterances and red for bad ones. This will allow the user to see quite intuitively if something is wrong without the interference of pop-up warnings hence preserving the flow of the recording. The participant can still be instructed to contact a data collector for help if the quality is consistently bad.

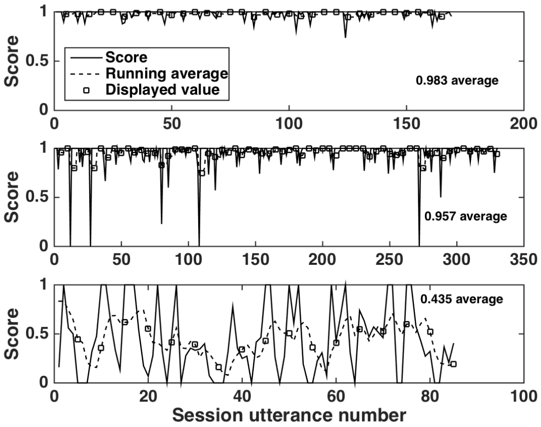


Figure 4: *The three panels show the phone-error score, the moving average and displayed value for three sessions from three different participants. The sessions' average scores are displayed in the panels.*

Figure 4 shows three sessions from the Málrómur data collection [10]. The solid lines show the inverted and capped PER score, the broken lines show a running 5 utterance average and the square points show the numbers reported back to the client. The three panels display a good, a medium-to-good and a bad session and the session averages are displayed. The average score over the entire session is given in each panel. The first session gives consistently good feedback to the participant even though the odd utterance scores below 0.9. The second session will get an occasional yellow warning and the last session is all in yellow or red.

4.1. System evaluation

The Málrómur database [10] was used to evaluate the forced alignment scoring. Figure 5 shows the average scores calculated over each of the 570 recording session of the entire database. The histogram shows that most sessions get good scores indicating a good fit with the aligner on average. There are however about 100 sessions that score less than 0.8 on average. These results give an idea about how the aligner works overall but to get more insight further analysis is needed.

4.2. Forced alignment performance

A listening test of 3000 recordings was carried out so as to discern the behaviour of the aligner further. Four annotators rated the recordings using Eyra's annotation module. A total of 2079 utterances were considered good by all annotators. This set of good recordings was copied three times but in each copy an error was created by modifying the text prompt by introducing a single substitution, deletion or insertion respectively. This created a total of 8018 utterances as some of the prompts were single words so they got deleted in the third copy.

Figure 6 shows a histogram of the aligner scores for all the four sets. The white bars show the distribution of the original 2079 utterances. Of those utterances, 1586 get a score of 1 and

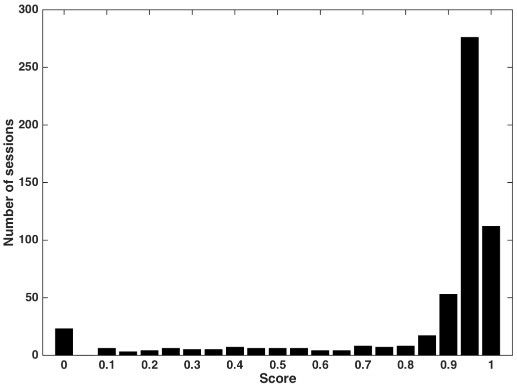


Figure 5: *Distribution of forced-alignment scores over 570 sessions.*

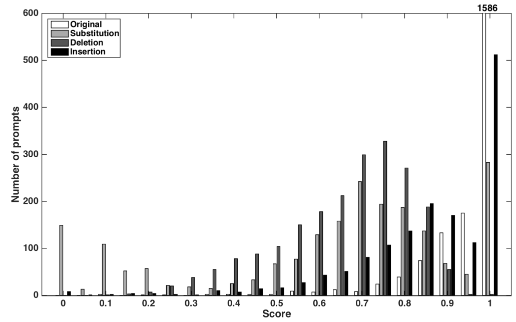


Figure 6: *Histograms of the scores given to each of the four prompt groups.*

less than 5% has a score beneath 0.8. This is a good result for the original set as a score of 1 is desirable but lower scores are to be avoided. For the other sets the reverse is true as we would like the aligner to catch those recordings with a low score. For the substitution set (light gray), 281 prompts got a score of 1 but most of the prompts (69%) from this set have a score below 0.9. For the deletion set (dark gray), the aligner worked very well. No utterance from that set got 1 or 0.95 and most of the utterances (over 80%) score 0.8 or lower. The insertion set (black) did not do so well. The number of utterances with a score of 1 was 509 and just under 70% of the set was above 0.8. The conclusion is that the aligner does very well in detecting deletions, reasonably well in detecting substitutions but some work needs to be done on insertions.

5. Eyra for data collection

The Málrómur database was collected [10] using Datahound [8] with the help of Goolge in 2011. The database is available at http://www.malfong.is with a Creative Commons BY 4.0 licence. The aim of the Eyra project is to make a data collection tool freely available but the design and implementation is built on the collaboration and the experience from the Málrómur database collection. Icelandic and Javanese speech data have been collected using the Eyra framework. The data was collected before the quality control module was implemented.

5.1. Icelandic data collected with Eyra

Data was collected over a period of two months at Reykjavik University. The database was recorded using ten smartphones running Android Lollipop. The back-end server was run on a laptop that connected to the smartphones over a wireless local area network. The collection therefore simulated an environment where internet access is intermittent or unavailable.

The total number of recordings is 32,929, 27,023 from the Málrómur prompt list and 5,906 from the tagged Icelandic corpus (MÍM) [12]. The total number of particpants were 203, of which 136 were male and 67 were female. The total duration of the recordings is 32.3 hours, an average duration of an utterance is 3.53 seconds. The data is made available on http://www.malfong.is with a Creative Commons BY 4.0 licence.

5.2. Javanese data collected with Eyra

The Javanese data was collected at Gadjah Mada University and Sanata Dharma University in Yogyakarta in May 2016. Unlike the Icelandic data collection, this effort was run in crowdsourcing mode, the back end server was run on Google Cloud and the participants used their own handsets. This was a good test for the Eyra system since there was little control over the client, connection was intermittent at times and the load on the server could not be controlled since arbitrary number of participants could be submitting utterances at any given time. The benefit of running the data collection in this way is, however, the potential speed in which data can be obtained. If one hundred participants turn up at the same time, they can all get started contributing to the database, hence shortening the collection period and/or increasing the quantity of utterances that can be collected.

The following actions deal with the challenges of open crowd-sourcing. First, as participants signed up, it was made sure that their handsets had the right version of Android set up. This mostly took care of client/handset issues but data collectors also monitored the participants well and got them other handsets if theirs didn't work. To deal with intermittent internet connection, a buffering scheme was developed for Eyra. Collection can take place on a disconnected handset, as long as there is a prompt list that is downloaded at the start of the session. The buffered utterances are then synchronized, either automatically, or manually using a sync monitoring feature where the number of uploaded and buffered prompts can be seen. The challenge of server overloading has to be dealt with by using normal load balancing practices.

The prompts for the Javanese collection were obtained from Wikipedia. For code-switching research, additional English prompts were added to the list and made up 9% of the total prompt list. A total of 160,266 prompts was recorded by 772 participants, 477 male, 292 female and 3 other gender. The total duration of the database is 257.7 hours, the average utterance is 5.79 seconds. The data will be made available on http://www.openslr.org with a Creative Commons BY 4.0 licence.

6. Acknowledgements

The authors would like to thank Oddur Kjartansson, Supheakmungkol Sarin, Maftuhah Ismail, and Martin Jansche at Google for their valuable advice and contribution. Also a special thanks is extended to Dedi Eko Yunanto Priyadi and Rita Anastasia Widiarti for helping with the Javanese data collection.

7. References

[1] M. Petursson, S. Klupfel, and J. Gudnason, "Eyra - speech data acquisition system for many languages," in *SLTU*, 2016.

[2] E. Charniak, D. Blaheta, N. Ge, K. Hall, J. Hale, and M. Johnson, "Bllip 1987-89 wsj corpus release 1," *Linguistic Data Consortium, Philadelphia*, vol. 36, 2000.

[3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2015.

[4] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[5] L. Yi, P. Fung, Y. Yongsheng, D. DiPersio, M. L. Glenn, S. M. Strassel, and C. Cieri, "A very large scale mandarin chinese broadcast collection for the gale program," in *Proceedings of the 7th Conference on Int. Language Resources and Evaluation (LREC'10), Valletta, Malta*, 2010.

[6] D. Graff *et al.*, "Fisher spanish speech," *LDC2010S01 DVD Philadelphia: Linguistic Data Consortium*, 2010.

[7] 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.

[8] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. J. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *INTERSPEECH*, 2010, pp. 1914–1917.

[9] N. J. De Vries, J. Badenhorst, M. H. Davel, E. Barnard, and A. De Waal, "Woefzela - an open-source platform for ASR data collection in the developing world," in *INTERSPEECH*, 2011.

[10] J. Gudnason, O. Kjartansson, J. Johannsson, E. Carstensdottir, H. H. Vilhjalmsson, H. Loftsson, S. Helgadottir, K. Johannsdottir, and E. Rognvaldsson, "Almannaromur: an open Icelandic speech corpus." in *SLTU*, 2012, pp. 80–83.

[11] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1, pp. 31–51, 2001.

[12] H. Loftsson, J. H. Yngvason, S. Helgadóttir, and E. Rögnvaldsson, "Developing a pos-tagged corpus using existing tools," in *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages LREC 2010, Valetta, Malta, 23 May 2010 Workshop programme*. Citeseer, 2010, p. 53.