Semi-automatic labeling of the UCU accents speech corpus

R. Orr, , M. Huijbregts, R. van Beek, L. Teunissen, K. Backhouse and D. A. van Leeuwen

University College Utrecht, Utrecht University, Utrecht, The Netherlands
CLST/CLS, Radboud University, Nijmegen, The Netherlands,
r.orr@uu.nl

Abstract

The annotation and labeling of speech tasks in large multitask speech corpora is a necessary part of preparing a corpus for distribution. This paper addresses three approaches to annotation and labeling, namely manual, semi automatic and automatic procedures for labeling the UCU Accent Project speech data, at multilingual multitask longitudinal speech corpus. Accuracy and minimal time investment are the priorities in assessing the efficacy of each procedure. While manual labeling based on aural and visual input should produce the most accurate results, this approach is prone to error because of its repetitive nature. A semi automatic event detection system requiring manual rejection of false alarms and location and labeling of misses provided the best results. A fully automatic system could not be applied to entire speech recordings because of the variety of tasks and genres. However, it could be used to annotate separate sentences within a specific task. Acoustic confidence measures can correctly detect sentences that do not match the text with an equal error rate of 3.3%

Keywords: manual annotation, acoustic event detection, automatic utterance segmentation, asr based task classification, efficacy

1. Introduction

The UCU Accent Project is a medium sized longitudinal multilingual multitask speech corpus, currently being collected at University College Utrecht (UCU), in the Netherlands. The student community is international and English is the lingua franca, providing a rich source of multilingual and multiaccented native and non-native speech data.

Three student cohorts are being recorded over a six year period and the final corpus size will be upwards of 1000 recordings, producing over 600 hours of speech. Each recording contains up to 12 speech tasks, including read texts of varying difficulty, monologues and dialogue. Speakers speak in both English and their L1.

For this paper, we have taken three approaches to segmenting and labelling the speech tasks – automatic, semi-automatic and manual – with a view to finding optimal efficacy and accuracy.

Manual labeling of speech data is still the most common form of labeling despite a large body of research on automatic methods. As observed by the many authors writing about manual and automatic labeling, for example, (Gut and Bayerl, 2004), (Campbell and Sagisaka, 1992), (Cole et al., 1994), (Truong and Trouvain, 2012), (François et al., 2012), there is much variability in accuracy and inter and intra annotator agreement, affecting the reliability and validity of the labels.

Annotators can be affected by different L1 backgrounds, different dialects or differences in perceptual acuity. Even where there is agreement on the item to be labeled, there is the problem of interpretation of the labeling strategy and of the application of coding (Cosi et al., 1991). Manual labeling of large or longitudinal corpora also lends itself to a changing team of annotators. Individual familiarity with the project, motivation and interest in the task, and the extent and type of previous training will also influence annotator judgement. While manual segmentation and labeling might be expected to produce accurate results in single instances, the tedious and repetitive nature of the task leaves it vulnerable to fatigue and simple human error.

With such a large number of speech recordings to be labeled for specific speech tasks, we look at how much of this work can be automated, and whether automatic procedures can provide sufficient accuracy while reducing time investment and vulnerability to error. Among the automatic systems we looked at are acoustic event detection, and task and utterance segmentation based on automatic speech recognition (ASR).

2. The UCU Accents Speech Corpus

The UCU Accent Project is a longitudinal corpus. Four consecutive cohorts of students are being recorded at five time points over the three years of their undergraduate study. The majority of the speakers have Dutch as their L1. Apart from Dutch speakers and native speakers of English, over 30 other native languages are represented.

Facilitators include faculty members from UCU and from department of Linguistics at the Utrecht University, masters students and undergraduate students. All facilitators use the speech data from the recordings for their own research and as such are invested in maintaining the quality of the recordings.

The speaker is recorded on eight different channels, via different microphones, while seated at a table in one of the faculty offices in the college. A more detailed description can be found in (Orr et al., 2011).

2.1. Speaker Tasks

The tasks are varied in difficulty and type, and include both native and non-native speech. The speaker is asked to perform between 9 and 12 speech tasks, including texts that are read aloud, monologues in both the native language and English (if English is not the native language), and a dialogue with the person facilitating the recordings. The texts are intended for a variety of research purposes, and include the production of sentences for use in intelligibility testing, sentences for use in prosody analysis, and texts containing shibboleths. The speaker tasks are listed in Table 1.

Table 1: *The speaker tasks for each recording.*

	Task Description
1	Speaker name, date and time
2	Extract from the Rainbow Passage (Fairbanks, 1960)
3	*Please Call Stella* *
4	*The Boy who Cried 'Wolf'* (Deterding, 2006)
5	Balanced sentence sets: intelligibility testing †
6	5 sentences for investigating rhythm (White and Mattys, 2007)
7	Extract: Declaration of Human Rights (L1) ††
8	Extract: Declaration of Human Rights (English)††
9	2 minute monologue L1
10	2 minute monologue English (formal topic)
11	2 minute monologue English (informal topic)
12	3 minute dialogue with the facilitator

* Speech Accent Archive, George Mason Univeristy http://accent.gmu.edu

† Task 5 refers to sentences from van Wijngaarden (Van Wijngaarden et al., 2002) in quantifying intelligibility of speech in noise for non-native listeners.

†† http://www.un.org/en/documents/udhr/

2.2. Variability of content in the speech files

The speech recordings contain quite some variability. Speech tasks were added as the corpus developed in order to make it comparable to other similar corpora. The order in which tasks were produced could also vary. Furthermore, the recordings were originally made with no clearly audible separation between tasks. After the first round of recordings, with concern for labeling the data, an audible separator was introduced between the tasks. The choice for an audible separator stems from the nature of the recording setup. It is not only a signal for separating tasks, but it is also a prompt for the participant, making clear when they should speak.

3. Automatic segmentation

3.1. Semi-automatic acoustic event detection

The presence of an audible separator between tasks suggested the use of acoustic event detection to locate the beginning of each task and label it, producing a table of timestamps and corresponding events, i.e. speech tasks.

For this, we used a lightweight fully-connected two-state hidden Markov model with Viterbi decoding, with states for 'speech' and 'event,' using the implementation developed in (van Leeuwen, 2005). The output probabilities are modelled using 16-component Gaussian Mixture Models. These were trained on five manually segmented files from the recorded speech data, so that in total, approximately 50 tokens for 'event' were used for training, and approximately 40 minutes of speech. The transition probabilities were manually defined, in order to have maximum control over the false alarm/miss trade-off and class sequence smoothing. The transition probability matrix was parametrised by a parameter p_{trans} and a prior over the states p_{bell} which represents the prior that in Automatic Speech Recognition is modeled by the language model. Effectively, in decoding every frame the transition matrix

$$\log \begin{pmatrix} 1-p_{\mathrm{trans}} & p_{\mathrm{trans}} \\ p_{\mathrm{trans}} & 1-p_{\mathrm{trans}} \end{pmatrix} + \log \begin{pmatrix} 1-p_{\mathrm{bell}} & p_{\mathrm{bell}} \\ 1-p_{\mathrm{bell}} & p_{\mathrm{bell}} \end{pmatrix} \quad (1)$$

is used to update the current maximum log likelihood path. We used a fixed parameter $\log p_{\mathrm{trans}} = -10$ for smoothing detection transitions, and controlled the false alarm/miss trade-off via p_{bell}. In manual checking of the detected events, it is much easier to dismiss a false alarm than to find a missing event. Hence we operated using a relatively high prior for the event.

As features, we used 13 MFCC[1] parameters plus deltas computed over 9 consecutive 25 ms frames taken at 10 ms intervals. We used the 'rastamat' implementation made available by Dan Ellis (Ellis, 2005) in GNU Octave (Eaton, 2002), using default HTK[2] compatible parameters. After feature extraction, an energy-based sound activity detection algorithm was employed, integrating energy in a 300–8000 Hz range. The criterion for frame selection was that the energy in this bandwidth was larger than 30 dB below the maximum energy in the recording. Non-sound frames were in effect removed from the modeling and the decoding, and later inserted in the time mark-up. This had the effect that sometimes the duration for the detected event appeared much longer than the actual event, because the Viterbi decoding would have spanned just over a larger chunk of silence. Since the detected events were only there to mark task boundaries, this was not considered a problem.

3.1.1. Evaluation

All detected events were manually checked. Events were played, and either accepted as instances of the

[1] Mel Frequency Cepstral Coefficients

[2] Hidden Markov Toolkit

audible separator, or rejected as false alarm with a single key-stroke. If the number of accepted events was too low, a manual audio and visual scan through the sound file was made to find missed events.

For 292 recordings, the number of events was 3530 (a mean of 12.1 per recording), where 381 false alarms were generated and 13 events were missed. The total recording time excluding events for these recordings was 97.0 hours, with a total duration of events of 3.24 hours. Note that these times include silence that may arbitrarily be attributed to 'event' or 'speech'. Hence we were operating at 3.9 FA/hour and 0.37 % miss rate. Listening to and judging the events took 67 seconds per recording, on average, whereby the total "wasted time" on the false alarms was about 1100 seconds. This should be compared to the roughly 800 seconds necessary to manually locate the missed events in the audio recordings.

3.2. ASR-based task classification

For the first 115 recordings, where no audible task separator was used, we developed an ASR-based task classification system, inspired by the work in (Moreno et al., 1998). Without transcriptions for the spontaneous speech parts of the recordings, we cannot align a recognition of the entire recording on the transcription. In particular, because we use a language model trained solely on the transcriptions, fragments of the spontaneous speech task might be incorrectly aligned to the tasks in the transcription.

Therefore, we aligned part of the recognition to the transcriptions of each separate task. For each task with N words in the transcription we took the first $\frac{3}{2}N$ words of the recognition and aligned the two texts. Then, we shifted the recognition with $\frac{1}{4}N$ words and again aligned $\frac{3}{2}N$ words with the transcription. We did this for each window of $\frac{3}{2}N$ words until we reached the end of the recognition. In the end, the window with the highest number of aligned words was considered to be correct. We determined the exact start and end time of the task by selecting the first and last word that was aligned correctly. By using only marginally larger windows than the transcription length, we minimised the probability that parts the recognition of other tasks are aligned to the transcription.

We used our in-house developed Automatic Speech Recognition (ASR) system "SHoUT" (Huijbregts et al., 2009) with models for English and Dutch, without any further adaptation to the acoustic models. The entire recording was first pre-segmented in utterances based on energy and a minimum pause criterion, which worked well for the clean recordings in this corpus. Recognition results of all utterances in both languages were used as hypothesis transcription in the alignment procedure described above.

3.2.1. Evaluation

We have evaluated this automatic classification approach on a test set of 48 recordings. Table 2 contains for each task the percentage of time that the recording was falsely classified as the particular task (false alarms) and the percentage of time that the audio was not classified as the particular task when it should have been (missed speech).

Table 2: Performance of the automatic task segmentation based on ASR. Task numbers refer to those in Table 1.

Task	False alarms (%)	Missed speech (%)
2	13	12
3	15	7
4	3	12
5	49	2

3.3. Manual segmentation

The files without the audible separator were also segmented manually. The files were opened in a waveform editor [3] to allow playback for the listener as well as visual inspection. When a new task was located, a label was placed at a time point just before that task. The label file information was stored in the same format as the files generated by the event detection system.

Files were segmented in batches of 20 to avoid fatigue. In practise, however, because of the difference in the length of the read texts, and because of the clear differences in energy patterns between read texts and spontaneous speech, locating the speech task was very easy for all tasks with the exception of the switch from monologue to dialogue in some cases.

The manual task segmentation results can be used as a reference for further work with the automatic ASR-based approach.

3.4. Automatic utterance identification

The goal of the automatic utterance segmentation was to find the identity of all the intelligibility sentences of Task 5 automatically. Variability in the content of the speech files makes a strict force-alignment over the entire task impossible. Additionally, each sentence was scored on pronunciation accuracy to allow automatic or semi-automatic selection of the sentences that were pronounced correctly. We used forced recognition of entire sentences and a separate phone loop recognition to compare the acoustic likelihood of each recognised sentence with the likelihood in the free phone loop.

[3] Audacity: `https://audacity.sourceforge.net`

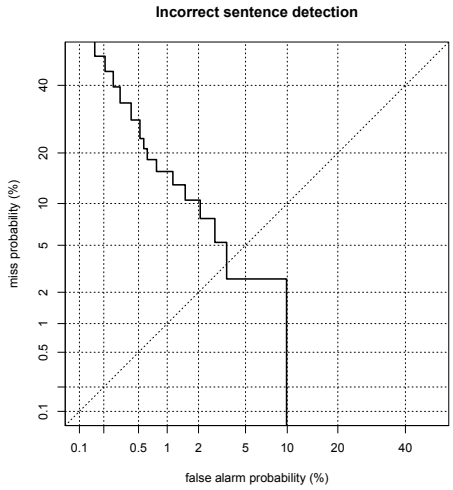

Figure 1: Detection Error Trade-off plot for the detection of errors in intelligibility testing sentences (Task 5) based on the acoustic confidence.

We created a pronunciation dictionary in which the words in each sentence were concatenated to one single 'word' in the dictionary. Utterances were segmented using a simple silence detector, which works because the intelligibility sentences are short (8–9 syllables) and spoken in a single effort, with pauses between consecutive sentences. The language model in this task was a simple uniform unigram model. In this manner, each complete sentence can be recognised irrespective of order they are pronounced.

3.4.1. Evaluation

The result of the forced sentence recognition was a list of start/end times of the sentences with acoustic confidence and the identity of the sentence. In order to evaluate the accuracy of the sentence recognition, we manually checked the recognition of 2010 sentences in 48 recordings.

A detection-error trade-off plot (Martin et al., 1997) is shown in Figure 1. This shows the trade-off between false positives and false negatives in error detection. The equal error rate is 3.3 %, which is a single-valued summary of the detection performance. For the sample of 2010 sentences, 38 were actually pronounced wrongly. At 10 % false alarms, there are no more mispronounced sentences missed.

From the 2010 sentences in the test set, 98.96% was correctly classified as correct. 84.21% was incorrectly classified as correct.

4. Discussion and Conclusion

Of the different methods that we looked at for the main task segmentation task, a combination of event detection and manual checking was the most efficient in terms of accuracy and time investment. The average 1.1 minutes per speech file to manually check the results of the acoustic event detection compares favourably to the ca. 4 minutes that it took to label a file manually. The low-miss-rate operating point we have chosen led to comparable costs of misses and false alarms, being 800 and 1100 seconds, respectively. This usually is an indication of a good choice of operating point. Furthermore, the manual labeling could take quite considerably longer where the listener had to search for the change from monologue to dialogue, or to search in the balanced sentence sets.

The ASR-based task segmentation was not sufficiently accurate for our recordings. We have a clear preference now for working with the event detection system, alongside manual checking. A semi-automatic system seems to be the most efficient way of segmenting the tasks in this corpus.

The ASR-based utterance segmentation works well for the intelligibility sentence sets in Task 5. The manual checking script used in evaluation (cf. Section 3.4.) will be integrated into our labeling work flow, similarly to the event detection system.

The next steps include applying speaker diarization to the dialogue sections of the recordings, as well as further optimising the ASR-based task segmentation. Whether the ASR-based approach or the event-detection approach proves more useful in the end, we will always have to run a manual check on the results before delivering the corpus for distribution.

5. References

Campbell, N. and Sagisaka, Y. (1992). Automatic annotation of speech corpora. In *Proc. SST*, pages 686–691.

Cole, R., Oshika, B. T., Noel, M., Lander, T., and Fanty, M. (1994). Labeler agreement in phonetic labeling of continuous speech. In *Proc. ICSLP*.

Cosi, P., Flavigna, D., and Omologo, M. A. (1991). Preliminary statistics evaluation of manual and automatic segmentation discrepancies. In *Proc. Eurospeech*, pages 693–696.

Deterding, D. (2006). The north wind versus a wolf: short texts for the description and measurement of english pronunciation. *Journal of the International Phonetic Association*, 36(2):187–196.

Eaton, J. W. (2002). *GNU Octave manual*. Network Theory Ltd. ISBN 0-9541617-2-6.

Ellis, D. P. W. (2005). PLP and RASTA (and MFCC, and inversion) in Matlab.

Fairbanks, G., (1960). *Voice and Articulation Drillbook*, chapter 5, page 127. Harper and Row.

François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., and Zweigenbaum, P. (2012). Manual corpus annotation: Giving meaning to the evaluation metrics. In *Proc. COLING*, pages 809–818.

Gut, U. and Bayerl, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Proc. Speech Prosody*, pages 565–568.

Huijbregts, M., Ordelman, R., Werff, L., and Jong, F. (2009). SHoUT, the University of Twente submission to the N-Best 2008 speech recognition evaluation for Dutch. In *Proc. Interspeech*, pages 2575–2578. ISCA.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proc. Eurospeech 1997*, pages 1895–1898, Rhodes, Greece.

Moreno, P. J., Joerg, C., Thong, J.-M. V., and Glickman, O. (1998). A recursive algorithm for the forced alignment of very long audio segments. In *Proc. ICSLP*, volume 8.

Orr, R., Quené, H., van Beek, R., Diefenbach, T., van Leeuwen, D., and Huijbregts, M. (2011). An international english speech corpus for longitudinal study of accent development. In *Proceedings of Interspeech*, pages 1889–1892, August.

Truong, K. P. and Trouvain, J. (2012). Laughter annotations in conversational speech corpora—possibilities and limitations for phonetic analysis. In *Proc. 4th Intl Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2012)*, pages 20–24.

van Leeuwen, D. A. (2005). The TNO speaker diarization system for NIST rich transcription evaluation 2005 for meeting data. In *Proc. Rich Transcription 2005 Spring Meeting Recognition Evaluation*, pages 84–92, Edinburgh. NIST.

Van Wijngaarden, S. J., Steeneken, H. J., and Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *J Acoust Soc Am.*, 111(4):1906–16, April.

White, L. and Mattys, S. (2007). Calibrating rhythm: first language and second language studies. *Journal of Phonetics*, 35(4):501–522.