LIGHTLY SUPERVISED LEARNING FROM A DAMAGED NATURAL SPEECH CORPUS

Charles Fox, Thomas Hain

Department of Computer Science, University of Sheffield, UK

ABSTRACT

Large corpora of transcribed speech are rare and expensive to acquire, but valuable for ASR systems. Of current research interest are corpora of natural speech, i.e. far-field recordings of multiple speakers in noisy environments. In the big data era there are many speech transcriptions collected for purposes other than ASR, which omit features required by typical ASR systems such as timing information. If we could recover training data from such 'found' corpora this would open up large new resources for ASR research. We present a case study for this type of data recovery – becoming known as 'lightly supervised learning' – for a highly damaged corpus called Family Life. We use a novel comparison of a parallel decode and forced audio alignment to iteratively select and grow good data. Family Life also has unusual data mislabelling problems which can be addressed by an integrated *tfidf* approach. These methods reduce WER on the corpus from 83.0 to 57.2. We also discuss a probabilistic loose string alignment approach which removes untranscribed 'icebreaker' speech.

1. INTRODUCTION

The UK Economic and Social Data Service (ESDS) contains around 5,000 data collections from social science studies from the 1960s to the present day, including raw audio interviews and transcripts. We work with one called Family Life[13] as a case study, showing how timing information can be recovered from a corpus where damaged and permuted text is wide-spread and long audio recordings do not come with any timing annotations. This opens up Family Life and similar corpora for use in social-, market- or other corpus-based analysis as well as turning them into resources for speech technology development - for example automatic speech recognition (ASR). Our aim is to retrieve good timing annotations on a word level for as much data as possible. We take a bootstrapping, iterative approach which alternates between forced audio alignments and training to grow the amount of data that can be aligned and improve trained models in the process. This paper tests a sequence of variations of increasing complexity on this strategy: first a basic form; then extending to handle Family Life's tape labelling permutations and an approach combining audio alignment with parallel biased decoding to select only good aligns for training.

Our method is closely related to other recent experiments in lightly supervised training. The field tends to use the same set of components – forced and loose alignments; text and audio alignments; scoring and iteration – in various architectures to approach the problem in different ways. [2] ran biased decodes, forced alignment between the decodes and transcripts, then loose audio alignment to the segmented decodes. [6] use manual annotation as a seed, then iteratively run biased decodes and trains directly from the decoder output. [8] run a biased decode then forced-align the audio to the decode. [15] use a loose audio alignment. Systems have mostly been developed on clean audio – lectures, broadcast news and audio-books – which lack the range of noise and tape labelling permutations of Family Life, so yielding better performance (lower word error rates) than we expect for Family Life. However our aim is not to recognise the audio especially well, but to achieve confidence in as many aligned hours as possible. In contrast to these systems, we deploy parallel forced audio aligns and decode, then compare them to select training data, and iterate. (Less related approaches have also used techniques such as SVMs[10], factor automata [7] and machine translation [5].)

2. THE ESDS FAMILY LIFE CORPUS

Family Life is a set of one-on-one interviews with elderly people describing life in the UK between 1900-1918, made in 1967[14, 13]. There were 452 interview subjects, and interviews last for around 3 hours. There total audio amounts to 1,354 hours, almost all is speech, all interviews are transcribed. Multiple transcribers used different conventions for fragments, dialect words, and non-speech annotations. The recordings were made with mono microphones on analogue tapes of various lengths, most commonly 45, 30, 15 and 60 minutes[1]. Not all tapes are filled with data: Some interviews use up whole tapes and continue immediately on the next one; others terminate the interview part way through a tape and begin on a new one, possibly on a different day. Interviews typically comprise 3 or 4 tapes (10 maximum), digitised to 44.1kHz stereo wav files.

Interviews are loosely structured over a set sequence of topic themes, such as school, religion, work. Questions are not prescribed but the corpus comes with an interviewer guide-sheet, giving an order in which topics are likely to occur. There are 44 named interviewers, conducting 8.2 ± 7.2 interviews each. A further 84 interviews are conducted by unknown interviewers. The 452 subjects are from 11 UK regions, with 40.8 ± 17.0 subjects per region. 230 subjects are female, 222 male, of ages 72 ± 7 years. An A-G social class is provided for each subject, with 64.5 ± 35.9 subjects in each. Other data provided includes rural/urban environment; marital status; job description and category; interview location. Being recorded in the 1960s with elderly subjects, the diversity of accents is notably greater than in modern recordings. Most subjects have strong regional accents, usually related to the interview region, where many have lived all their lives; the interviewers are 1960s sociologists, many having strong 'received pronunciation"(RP) accents. A few interviews are with non-native English speakers, sometimes in their native language (e.g. Welsh) but transcribed in English.

2.1. Data issues

Many tapes begin with 30s-2mins of non-transcribed 'icebreaker' conversations, in which the interviewer tests the tape recorder, sets the volume, and chats casually with the subject. Similarly, there are

[1] We use 'tape' to mean audio from one side of a two-sided cassette.

many untranscribed utterances that we call 'interjections', including prompts and repetitions by interviewers, and short remarks irrelevant to the interview topic by the interviewee. Speech is far-field, the tape recorder usually being placed on a table between interviewer and subject; common background noises include traffic; ticking clocks; furniture creaks. Most seriously, errors have been made in the tape labelling and/or digitisation. For many interviews, one or more tapes is either missing, permuted within the same interview, or permuted with tapes across interviews. Around 20 interviews are missing all transcripts or audio. As transcriptions were not made for ASR work, they contain *no timing information* and do *not even show the locations of the tape boundaries*. In places, often near starts of tapes, digitisation errors have resulted in tape noises and out-of-order audio, for example playing the tape for 30s then rewinding or fast forwarding it – putting the cued audio into the digitiser.

2.2. Processing

After text normalisation, the corpus contains 560598 script-style, alternating lines between interviewer and subject. There are 11.4M word tokens (59Mb text file), of 51,000 distinct words. Perplexity of the corpus under a generic language model [1] is 144. 49% of words only occur once, possibly due to many place names. An indication of topics is given by the top nouns: YOU, ME, MOTHER, SCHOOL, FATHER, PEOPLE, WORK, HOUSE, DAYS, CHILDREN, THINGS, SUNDAY, FAMILY, PARENTS, CHURCH, FRIENDS, JOB, MONEY, STREET, CLASS, BED, ROOM, BOYS, SISTER, TEA, CLOTHES, WAR, LIFE, SHOP. Common non-standard words include Scottish contractions (DIDNAE, COULDNAE) and terms and symbols for old English money.

Audio was down-sampled to 16kHz, 16bit and encoded as 12 PLP coefficients and c_0 [3]. First and second order derivatives were added forming a 39 dimensional feature vector. No further normalisation was applied. Using the meta-data, a test pool of 30 interviews was selected automatically, having similar region, age, and gender distribution to the full corpus. Two test sets consisting of different 5 minute extracts from each test pool member were constructed. They were manually checked to remove permuted or otherwise damaged interviews, and manually annotated with ground truth start and end times. The remaining interviews form the training pool. Each step of our process consists of creating a new time-annotated data set then training new acoustic models. Standard 3-state left-to-right hidden Markov Models (HMMs) are state-clustered phonetic decision tree ties state with 16-component GMMs to model output probabilities. Training follows a standard HTK mixup procedure [17]. Word error rates (WERs) are obtained with NIST sclite [9]. Decodes are based on HTK HDecode with per-interview biased language models built from the transcripts with the SRI language model toolkit[11]. Each alignment referred to is followed by a resegmentation step, cutting into short utterances using silence detection.

3. BASELINE ALIGNMENT AND TRAINING

Initial experiments were performed using a basic iterative bootstrapping approach as is commonly used to obtain acoustic models in a new domain. This approach serves as baseline and illustrates the complexity of the task at hand. Initially we have no knowledge of timing information, and thus are unable to split the data into manageable chunks (the length of interviews is up to 3 hours!). Thus a rough alignment of the entire corpus is performed using an initial model set, M0, trained on 177 hours of timing-annotated general meetings audio, recorded with close-talking microphones [1]. Once a rough alignment was obtained, a first Family Life model, M1, was trained. M1 is then used for alignment again, and the new alignment is used to train a model M2.

In all of these iterations, we attempt to align all 452 interviews using no prior segmentation – each is a single, 3 hour long segment – and alignment often fails due to aforementined data issues. Results are shown in table 1. Successful alignment, even with tight pruning, does not imply that the aligned times are actually correct. A slightly wider beam (400) than usual was chosen to allow alignments to recover from gross mismatches between labels and acoustics. Such settings are normal for a standard corpus. It is also important to note that there is no guarantee that the same data aligning in iteration 1 does align again in the second round, though the size of of the training sets tends to grow overall tables columns ex,ai,ah). As shown in the table, the better matching in-domain models allow this basic scheme was able to grow the aligned data at each iteration, to 217 for M2. However it was clear from a 94.4 test WERs that the training is going wrong: incorrectly aligned interviews are being admitted to training, and encouraging more incorrect alignments. Despite more aligns, the test scores are worse than the 83.0 of M0, showing that this basic iteration scheme is insufficient for this corpus.

4. HANDLING TAPE PERMUTATIONS AND BAD ALIGNS

An unusual cause of WERs in Family Life is its tape labelling errors. Given the size of the corpus a fully automated method for correcting such errors is desirable. There are three types of labelling error. First, two or more tapes within the same interview may be transposed or otherwise permuted (imagine the interviewer writing 'subject 453 side 1' and 'subject 453 side 4' on the wrong cassette sides, or the digitisers making similar errors). Second, tapes *between* interviews may be permuted (e.g. swapping 'subject 68 side 3' with 'subject 96 side 2'). Third, sections of transcripts and audio are missing. A more generic WER cause is failure of the basic iteration scheme to converge on correct alignments, resulting in training on bad data. We present methods for handling the most common, intra-interview, permutation case, and for filtering out bad alignments.

4.1. Tape label permutations

We assume audio and transcripts are complete and free of spurious data, i.e. there exists a permutation of the tapes resulting in good alignment. No timing information is available as initial forced alignments are unreliable, so the location of the tape boundaries in the transcript is also unknown and needs to be estimated automatically.

For each permutation of the N tapes, the transcript is split into N chunks corresponding to the tape lengths in such a way that they have approximately the same lengths as the corresponding tapes under that permutation. This is done by ensuring they have the same ratios of number of words as the tape sides have of audio duration. Each permutation of the transcript is then scored against a decoding of the whole interview, here using the M0 model. To score, each tape's decoding is treated a query, q, in a term-frequency/inverse-document-frequency [4] model,

$$tfidf(q, i, \{doc_j\}_j) = \sum_{w \in q} \frac{f(w,q) f(w, doc_i)}{f(w, \{doc_j\}_j)}, \qquad (1)$$

where $f(w, d)$ is the frequency of word w in document or set of documents d, and doc_i is the ith segment of the hypothesised permuted

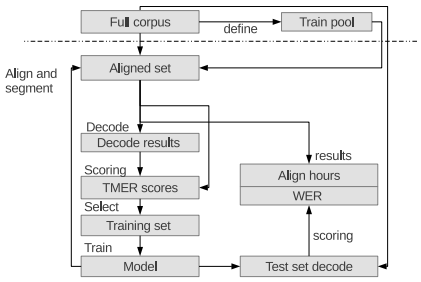

Fig. 1. System architecture for showing iterations of parallel alignment and decoding; and training. Each iteration creates an enlarged training set and better model.

transcript. A score sc is then assigned to a complete permutation,

$$sc(perm) = \sum_i tfidf(q_i, perm(i), \{doc_j\}_j), \qquad (2)$$

where $perm(i)$ is the permutation index for the ith segment of the permutation, i.e. we sum scores for each decode segment in the permutation. The best permutation for each interview is retained.

4.2. Filtering of mismatched audio

Another source of error occurs when alignments are very poor and HMM training learns from inconsistent or mismatched data. We saw that M1 and M2 align increasing quantities of data but with poor WERs suggesting that training on either aligns or decodes from these models – as in previous systems – is counterproductive. Under normal circumstances HMM training is known to recover well from such situations, but if the amount of badly aligned data is too large, significant modelling errors occur and compound. From a data selection perspective the outcome of each interview's alignment is binary: it either survives or it fails. However, we may look for subregions *within an interview* that appear to be well- or badly aligned. For those that align mismatched data can be filtered using output from recognition of the training set. A full decode of all interviews is obtained alongside the forced alignment. This is followed by standard DP alignment between the two time-segmented sequences. Rather than using word error rates of whole segments, we define a smoothed, per-word *temporal matching error rate, TMER*, by

$$TMER[t] = \frac{1}{N} \sum_{t'=t-N}^{t} (C[t'] - S[t'] - D[t'] - I[t']), \qquad (3)$$

where C,S,D and I are Boolean functions indicating correct, substitutions, deletions and insertions, and t ranges over the union of decoded and transcript words from the scoring. On inspection, as in fig. 2, the TMERs typically fall into two regimes – aligning and not-aligning – for most interviews, with a typical interview beginning with successful alignment, then losing tracking for a while, then perhaps regaining it again. The TMER in the non-aligned sections is expected to be roughly constant, reflecting the error rate of matching random words to audio, while any parts that align will be significantly better than this, reflecting the local word error rate. It was found to be sufficient to simply filter segments on the basis of thresholds on TMER values. Segments are assumed to be correctly aligned when $TMER[t] < -0.75$, with $N = 100$.

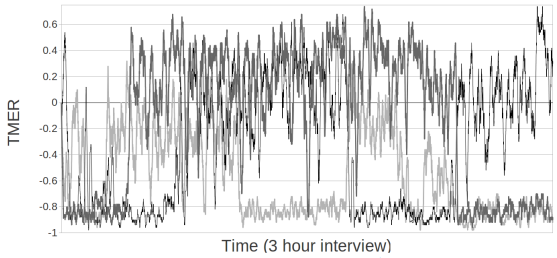

Fig. 2. Typical TMER scores for three interviews, each plotted over three hours.

A new new training set can be constructed based on TMER filtering, discarding all bad permutations and badly aligned segments. This leads to an extended version of the iterative bootstrap training cycle used in the previous section, and show in fig. 1. Note that at each iteration the best model so far is used to both forced-align and decode the current best training set. The outputs of these processes are synchronised by TMER to define a new, correct-align training set, which then is used to train the iteration acoustic model.

4.3. Joint icebreaker removal and permutation repair

The tdidf approach to permutation detection makes a big approximation in estimating the tape boundaries in the reference text: it assumes that the ratios of transcript words are equal to the ratios of the audio tape lengths. This is unfounded when, for example, full tapes are not used in interviews and there are many minutes of silence at tape ends, or when long icebreakers or interjections occur in some tapes. Icebreakers and interjections also reduce alignment quality, as they are forced to align against transcripts that do not include them.

A more complex approach to permutation repair avoiding tape boundary approximation, which could also handle icebreakers and interjections as a side effect, is given by considering a further combination of elements found in the reviewed previous systems: decoding, loose alignment, and string alignment. Here we first run a full decoding of each interview, then align the decode text to the transcript text to produce training data. String alignment is much faster than audio alignment, and can be performed with loose string HMMs that model permutation, icebreakers and interjections.

For each current transcription position, beginning at the transcript start, we built a loose HMM based on the transitions though the reference text from that position, but with probabilities of word insertions, deletion and substitution taken directly from previous NIST scoring runs. Incorrect word observation distributions were modelled by the first order corpus word distribution. The first state of each HMM is modelled as an icebreaker state, having a large self-transition (using an exponential prior with mean one minute icebreaker length) and generating all words from the corpus distribution. We then ran Viterbi alignment on each remaining tape decode, and selected the best tape and alignment. This provides a greedy $O(N^2)$ search over permutations, aligning one tape decode at a time and finding precise tape boundary in the reference text at each step. Some example aligned text from the model is shown in fig. 3.

4.4. Experiments

Interviews with more than six tapes and various data glitches were removed due to the computational load of $O(N!)$ tdidf search, so

we ran on 287 interviews. First we tested the accuracy of the tdidf algorithm itself. 74 interviews were reported to have incorrect permutations (25%), 21 of which were human checked. Of these, 5 were inter-interview permutations where the algorithm gave correct permutations; 1 was an inter-interview permutation reported as an incorrect permutation; 8 were between-interview swaps; 2 were interviews in Welsh language but transcribed in English. 5 were false positives having high levels of tape and environmental noise and/or strong dialects. (The algorithm may fail more when when there are many tapes: if we further exclude the six-tape interviews we have 60 of 272 problems, or 22%). Of another human checked 21 interviews reported with no permutation problems, 2 were false negatives (one in Welsh and the other missing 10% of the audio at the end).

Applying *tfidf* permutation removal and TMER alignment filtering in three iterations produced models M3-M5 whose results are shown in table 1. These indicate that by removing both bad permutations and badly aligned segments we can construct improved models (M5 WER of 63.6, vs M0 WER of 83.0) and alignments (165h from M5 vs 104h from M0) of more of the corpus. This is in contrast to the basic M1,M2 iterations, which aligned more (217h) data but apparently converging incorrectly as evidenced by their large WERs. However the iterative process appears to quickly reach a new local minimum state where little further gains are available from iterations (WER stuck at around 63), so a new technique is needed at this stage. To escape from this local minimum, we tried to restore, rather than discard, the badly permuted interviews. The best $tfidf$ score comes with an estimated (as the tape boundaries in the reference text are only approximate) permutation needed to re-order the tapes correctly. We created a new, larger training set including such repaired permutation, and performed two further iterations to produce the M6 and M7 models whose results are shown in table 1. Perhaps due to the crude approximation of tape boundaries, M6 does not align a larger quantity of data than its predecessor M5 422h vs 424h), though the aligns are of higher quality as evidenced by the increase in post-TMER data size (tsh of 273 vs 269) and improved WER (57.2 vs 63.6).

Our present implementation of string based alignment removes only icebreakers, not interjections, and has an implicit flat prior over possible permutations. Beginning with decodes and NIST probabilities from the M5 model (to enable comparison with the M6 tfidf method) we aligned 426 hours of the training set and obtained a WER of 76.1, a comparable performance to the M5 model on which its decodes were based. This shows that the algorithm is working but not yet very useful. Inspection of the aligns suggests that the main failure case occurs for very damaged interviews: if string align tracking is lost just once then both alignment and permutation estimation are irrecoverably destroyed, leading to worse training data that the tdidf method. This could perhaps be corrected by using an informative prior on permutations, so that when aligns give weak likelihoods, the default ordering would be used (or the interviews removed altogether, as for failed audio aligns in the tfidf/TMER version.)

5. CONCLUSIONS

We have demonstrated two new methods to repair damaged transcriptions having tape label permutations and no timing information, and produced good alignments for a useful subset (273 hours) of the natural speech Family Life corpus. We saw that simple forced audio align and train iterations are not sufficient for highly damaged corpora such as Family Life, but that by introducing a TMER-based comparison between parallel decodes and forced-aligns along with tfidf based permutation restoration, progress can be made.

Once a good model has been created from this process, the audio alignment may be dispensed with and a loose string alignment between decodes and transcripts used for lightly supervised training. We have showed an implementation of this with a WER comparable to the best audio/TMER approach, and seen that it shows potential for further improvements by removing interjections and using informative permutation priors.

Our methods use similar building blocks to other lightly supervised systems: mixtures of forced and loose alignment; string and audio alignments; data selection; in new combinations. The field has not yet matured to running comparison studies between methods and corpora but we suggest this as an important future topic.

Table 1. *Results. m=model; ai=number of aligned interviews; ex=number of exits (interviews that previously aligned and now fail); ah=aligned hours; adh=hours successfully aligned and decoded; tsh=TMER-selected hours; I,D=insertion,deletion rates.*

m	ai	ex	ah	adh	tsh	WER	I	D
M0	104	-	286	-	(95)	83.0	9.9	15.5
M1	89	15	285	-	-	95.7	5.7	23.7
M2	217	1	559	-	-	94.4	3.9	24.1
M3	133	9	333	206	150	64.0	15.3	12.6
M4	153	5	385	231	158	63.1	17.0	11.0
M5	165	0	424	263	269	63.6	17.5	10.6
M6	167	0	422	406	273	59.3	9.7	18.5
M7	-	-	-	-	-	57.2	18.6	9.9

Fig. 3. String alignment. Green=correct, red=incorrect matches. The vertical line in an interjection or icebreaker, which could be removed on detection to match the transcript.

6. ACKNOWLEDGEMENTS

Thanks to the ESDS hosted by the University of Essex in Colchester, UK for access to the data. We benefitted from work by Vincent Wan for initial investigations. Funded by EPSRC EP/I031022/1.

7. REFERENCES

[1] Thomas Hain, Lukas Burget, John Dines, Philip N Garner, Asmaa el Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan. The AMIDA 2009 Meeting Transcription System. In *Interspeech'10*, pages 358–361, 2010.

[2] Timothy J. Hazen. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *Proceedings of Interspeech*, 2006.

[3] Hynek Hermansky. Perceptual linear prediction (PLP) analysis of speech. 87(4):1738–1752, April 1990.

[4] Karen Spark Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[5] Tatsuya Kawahara, Masato Mimura, and Yuya Akita. Language model transformation applied to lightly supervised training of acoustic model for congress meetings. In *Proc. ICASSP*, 2009.

[6] Jean-Luc Gauvain Lori Lamel and Gilles Adda. Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language*, 16:115–129, 2002.

[7] Pedro J. Moreno and Christopher Alberti. A factor automaton approach for the forced alignment of long speech recordings. In *Proceedings of ICASSP*, 2009.

[8] Sabine Buchholz Norbert Braunschweiler, M.J.F. Gales. Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *Proceedings of Interspeech*, 2010.

[9] National Institute of Standards and Technology (NIST). *Speech Recognition Scoring Toolkit (SCTK) Version 2.4.0.* web resource: http://www.itl.nist.gov/iad/mig/tools,, 2010.

[10] Kengo Ohta, Masatoshi Tsuchiya, and Seiichi Nakagawa. Detection of Precisely Transcribed Parts from Inexact Transcribed Corpus. In *Proc. ASRU*, 2011.

[11] Andreas Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. Spoken Language Processing*, 2002.

[12] Annika Hamalainen Antonio Calado Miguel Sales Dias Daniela Braga Thomas Pellegrini, Isabel Trancoso. Impact of age in asr for the elderly: Preliminary experiments in european portuguese. In *Advances in Speech and Language Technologies for Iberian Languages Communications in Computer and Information Science, pp 139-147*, 2012.

[13] P. Thompson and T Lummis. Family life and work experience before 1918, 1870-1973. *Economic and Social Data Service*, SN2000, 2002.

[14] Paul Richard Thompson. *The Edwardians: the remaking of British society*. Psychology Press, 1992.

[15] Anand Venkataraman, Andreas Stolcke, Wen Wang, Dimitra Vergyri, Jing Zheng, and Venkata Ramana Rao Gadde. An efficient repair procedure for quick transcriptions. In *Proc. Interspeech*. ISCA, 2004.

[16] Ravichander Vipperla. *Automatic speech recognition for ageing voices*. PhD Thesis, Edinburgh University, 2011.

[17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, XA Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al. The HTK book. 2006.