



DERIVATION OF A LARGE SPEECH AND NATURAL LANGUAGE DATABASE THROUGH ALIGNMENT OF COURT RECORDINGS AND THEIR TRANSCRIPTS

P.E.Kenne, Hamish Percy
University of Canberra
PO Box 1 Belconnen ACT 2616 Australia

Mary O'Kane
University of Adelaide 5005 Australia

ABSTRACT

A major difficulty for both speech recognition systems and natural language systems is the large effort required to port such systems to a new application. Both speech and NL systems require large amounts of training data. The data collection and annotation is generally a labour-intensive activity.

All court proceedings in Australia are recorded, and transcripts are produced for over 95% of them. The recordings, together with the transcripts, provide a rich source of data for speech and NL training. The court recordings are examples of spontaneous speech. Training using spontaneous speech (as opposed to read speech) can significantly improve performance for recognising spontaneous speech [1].

A major difficulty in using these data to derive a speech recognition training database is that the transcripts are not in any way time-aligned with the audio data. We describe how we are deriving a large speech recogniser training database from Australian court recordings in a semi-automatic manner through aligning the court recordings and their transcripts using a successive refinement bootstrap procedure which relies particularly on speaker-dependent word-spotting of common words.

BACKGROUND

The origin of this project was an enquiry from Auscript, the court reporting service of the Commonwealth of Australia, if it would be possible to use automatic speech recognition techniques to produce court transcripts. At present, all court proceedings are produced manually by two main methods: court reporters take down the proceedings using shorthand, and then dictate their transcripts to typists. (A shorthand reporter would generally take shorthand in a 15 minute block before being replaced by another.) The second method involves recording (using audio tape) the proceedings and having teams of typists produce the transcript. Partial automation was introduced by having the shorthand reporters dictate their transcript to a speaker-dependent speech recognition system (e.g. such as produced by Dragon Systems, Kurzweil, Verbex, etc.).

Full automation of the process of producing court transcripts requires the full sophistication of a large vocabulary speaker-independent speech recognition

system in conjunction with a suitably sophisticated natural language system (to handle the editing requirements and conventions of Auscript). Achieving such an integrated system remains an open research problem, however, considerably more automation than is used at present is potentially achievable by using extensions of existing techniques.

In this paper, we concentrate on techniques to derive a large speech database through the alignment of court recordings with their transcripts. Long-term, this database will provide a valuable resource for the training of any speaker-independent speech recognisers for Australian English. However the process of deriving the database provides a path to automatic transcription of a very large percentage of court transcripts. The database also provides the basis for a natural language database for language encountered in the courts.

CHARACTERISTICS OF THE SPEECH DATA AND CHARACTERISTICS OF THE COURT ENVIRONMENT

Courts are located in many different styles of rooms, and the acoustic conditions are highly variable. In addition, there has been no attempt to standardise the recording equipment until recently, when the decision was made to use DAT technology and to standardise the type of microphone used. Microphones are positioned in front of the judge; the lawyers for each side have a separate microphone; and there is a microphone located in the witness box. Each of these parties is recorded on a separate track. Even with high-quality noise-cancelling microphones, the head and body movements of a lawyer, the judge or a witness results in variable quality recording. In a small number of cases (currently estimated to be 0.5%) the recording is untranscribable.

A typical court vocabulary is about 15,000 distinct words, and of these, about 500 words account for 80% of all words spoken in these cases. In any given case, there are a number of high frequency case-specific words, and also a number of high-frequency words specific to a particular stage of the court proceedings. For example, in the first day of a case involving a dispute between a number of unions over which union should represent a group of workers, the words "application" and "crossapplication" occur frequently, but occur far less frequently on subsequent days. There are also a number of words and phrases which could be labelled as court courtesy phrases (e.g. "your honour", "may it please your

honour", "our respectful submission", "my learned friend") which occur with high frequency.

The types of speakers can be generally characterised as a small number of speakers who say a lot (the lawyers and judge), and a larger number of speakers who say relatively little (the witnesses). The speech of the lawyers and judge is often relatively long statements followed by a question, and the witnesses often answer with one or two sentences, or in many cases, with a single word. For example, in the case referred to above, a five-day hearing, the lawyers account for 64.5% of the utterances, the witnesses for 25.5% and the judge for 10%. The size of the lexicon varies according to the category of speaker. Similar observations may be made about most court cases dealt with by the Commonwealth. Table 1 below gives average details for several cases in one jurisdiction.

	Vocabulary (words)	Total words	Avg. length
Case 1			
Judge	1679	17122	15.3
Lawyers	4475	110517	22.4
Witnesses	3203	43655	12.3
Case 2			
Judge	1162	8369	12.8
Lawyers	6358	200574	24.2
Witnesses	5496	123370	16.3
Case 3			
Judge	611	3201	10.2
Lawyers	4506	68106	26.9
Witnesses	3559	45066	19.4
Case 4			
Judge	6502	112519	14.8
Lawyers	18288	1369465	28.8
Witnesses	16257	970819	23.5
All cases			
Judge	7126	141211	14.6
Lawyers	19981	1748662	24.4
Witnesses	18355	1182910	21.6
Totals	26634	3072783	24.0

Table 1: Lexicon size, total words spoken, average utterance length.

It is well known that the grammar describing spoken language is different from that describing written language [2]. However, the language used in typical interactions in the court room differs from casual conversational speech, partly due to the adversarial nature of proceedings. For example, a typical monologue from a lawyer is:

No, nor am I, your Honour, but what I am saying is that what his application - what this application so-called does is this: there is a matter before the commission that is allocated. It is allocated for some six months, and it deals with certain factual issues. What this cross application does is seek to broaden them out to include the whole of the civil and mechanical engineering sectors, the whole of the brass and non-copper ferrous - I'm sorry, brass, copper and non-ferrous metals industry, and the like. Now, in the brass, copper and non-ferrous metals industry, the . . . inaudible. . . trades on my understanding have little or no interest, the XXXXX have little or no interest.

In summary, the data from each court case provides a lot of speech data from a small number of speakers and a small amount of speech data from each of a large number of speakers. A small number of distinct words accounts for a high percentage of the transcript for all these speakers.

PERFECT TRANSCRIPTION OR ADEQUATE TRANSCRIPTION?

Analyses of the characteristics of court case data led us to the conclusion that speaker-dependent word-spotting of the commonly-occurring words and phrases used in court dialogue would lead to a high total-word transcription. Accordingly, we have used this observation as a guide to the approach of deriving a database from the court recordings and their transcripts. The database we derive should, as a practical intermediate goal to full transcription, allow us to train speaker-dependent recognisers (or wordspotters) which will recognise about 90% of all the words encountered in court transcript in any particular jurisdiction. (Generally judges and lawyers work only in one specialist jurisdiction.)

With this intermediate goal in mind, we have developed an approach to derivation of a large speech database which has as its essence an iterative bootstrapping procedure which is a mixture of automatic and machine-assisted human inspection phases. As the database increases in size (both in terms of the number of examples of each word and in the number of words included), the amount of human inspection and adjustment lessens and the usefulness of the database for training of general-purpose speech recognisers increases. The automatic phases of the procedure include the use of an initial alignment algorithm which provides the first approximation to an alignment of recording and text and the use of speaker-dependent wordspotters. The final database to be produced will be a combination of a large number of speaker-specific databases. Each of these databases can be used to train speaker-dependent recognisers. The overall procedure for deriving the final database is given in Table 2. We will now proceed to discuss significant features of the procedure.

It should be noted that the database we are deriving here is marked up at the word level not the phoneme level. However several procedures using dynamic programming techniques for phoneme marking once word level marking has occurred have been described [3], [4]. If one wished to derive phoneme-level marking, such procedures could be used.

THE PROCEDURE FOR ESTABLISHING THE DATABASE

Pre-alignment processing

As can be seen from Table 2, an amount of preprocessing is needed before any attempt at alignment of recording and transcript is undertaken. For each court case the speech is broken into consolidated blocks for each speaker, possible because of the separate tracks and the time sequencing

Establishing speaker files

Separate court transcript into consolidated blocks for each speaker retaining information of when the speaker commences and finishes each small sub-speech.

Note speakers for whom there are more than 5000 words of transcript. (Class 1 speakers)

Establish descriptor files for each speaker. Include speaker type (judge, lawyer, witness) and speaker sex information.

Text analysis

Carry out statistical analysis of complete transcript (lists of common words, word pairs, phrases, etc.). Calculate statistical variations between different days of the full court hearing.

Produce rough phonetic transcript of speech for each speaker by dictionary look-up.

Count total number of phonemes per speaker. Multiply by transcription-miss factor to obtain estimated total number of phonemes for each speaker [etnph].

Speech recording analysis

Run silence detector (silence > 40msec) over consolidated block for each speaker. This produces silence/speech labels for speech recording.

Sum lengths of non-silence parts of consolidated block for each speaker to give total speech length [tsl].

Find average phoneme length [avphlen] for each speaker. $avphlen = tsl / etnph$

Initial alignment algorithm

Produce initial time-alignment of transcript and recording. [This algorithm first finds rough approximation to length of each word for each speaker (no. of phonemes per word multiplied by avphlen) and then moves through the recording estimating the start and end of each word by using the estimated length of each word and taking into account the silence/non-silence mark-up of the recording and any annotations on the transcript which indicate major pauses.]

Training material for first-pass wordspotters

An annotation assistance tool is used to refine the results of the initial alignment algorithm. This detailed annotation is carried out on the first 1000 words of the speech of each Class 1 speaker. This information is used to derive training data for speaker-dependent wordspotters designed to recognise the 50 most-frequent words, the 20 most-frequent word pairs and the 5 most-frequent word triples.

Using wordspotters and phonetic rules to refine alignment automatically

Three or more speaker-dependent wordspotters and a number of sex-dependent but otherwise speaker-independent rule-based phone-event spotters are run over the next 1000 words of each speaker block.

Check consistency of results from wordspotters and rules, eliminating likely false fires and signalling likely transcription errors.

The annotation assistance tool is used to inspect these results and refine this alignment manually.

Bootstrapping to increased alignment efficiency

Use two-thousand word accurate alignment passage to re-train the wordspotters and increase the number of words to be spotted.

Run new spotters over the next block of speech; run the consistency algorithm; inspect results and refine manually until the whole of the speaker block is time-aligned.

This whole process is run for each Class 1 speaker.

Pooling of training data to get speaker-independent spotting

The results of the all the alignments for each speaker are then pooled and speaker-independent wordspotting training is carried out.

The speaker-independent wordspotter is run over all the speech from non-Class 1 speakers. Speech for these speakers has already been early aligned.

Table 2: Overall speech database production procedure

which can be derived from the transcript. Simple descriptor files are established for each speaker; these are augmented as the procedure progresses overall.

A text analysis phase is then commenced which includes statistical analysis to derive the most common words and phrases and the derivation of a rough phonetic transcription of the speech for each speaker by dictionary look-up at the word level. This can be used to estimate the total number of phonemes spoken by each speaker. Two factors complicate the estimation of total number of phonemes per speaker.

These are the inaudible pieces of recording referred to above and the fact that, due to Auscript editorial policy, the transcripts are not a completely faithful representation (neither at the word nor the phonetic level) of the audio. For example, repetitions such as "yes yes yes" are transcribed as "yes", and other effects such as stutters are not transcribed. So the original phoneme count per speaker is multiplied by a transcription-miss factor to achieve a better estimate.

Speech processing then begins with the application of a silence detector over the consolidated speech block

for each speaker. Non-silence sections are tagged and the total non-silence for each speaker is found by summing the lengths of these tagged sections. The average phoneme length per speaker ('avphlen') is found by dividing the total length of speech for each speaker by the estimate of the total number of phonemes. For the speakers incorporated in the database so far, 'avphlen' agrees well with figures reported elsewhere for English phoneme length [5].

Initial alignment algorithm

One of the crucial speed factors in deriving the database is the initial alignment algorithm. This algorithm first finds a rough approximation to length of each word for each speaker (no. of phonemes per word multiplied by 'avphlen') and then moves through the recording estimating the start and end of each word by using the estimated length of each word and taking into account the silence/non-silence mark-up of the recording and any annotations on the transcript which indicate major pauses.

This algorithm gives an average coverage of the total text of 72.6% with the average error in the start and finish times for each word being 58.4 msec and 74.0 msec. The algorithm does not cope well with short words but is a good guide to the approximate locations of long words. See Table 3 for an example.

Actual (ms)	Utterance	Predicted (ms)	%
start finish		start finish	cover
5329 5673	takes	5329 5662	96.8
5673 5733	the	5662 5828	100.0
5733 6077	view	5828 6077	72.4
	silence		
6605 6764	that	6605 6854	100.0
6764 7475	eligibility	6854 7768	87.2
	silence		
7560 8000	rules	7560 7892	75.6
	silence		
8330 8481	are	8330 8413	55.1
8481 8575	to	8413 8579	100.0
8575 8640	be	8579 8746	93.6
	silence		
8691 9021	altered	8691 9189	100.0
9021 9414	under	9189 9522	57.2
9414 9871	section	9522 9937	76.5
9871 10999	118a	9937 11349	94.1
	silence		
11922 12070	and	11922 12171	100.0
12070 12151	in	12171 12337	0.0
	silence		
12192 12386	our	12192 12358	85.6
12386 13031	respectful	12358 13105	100.0
13031 13698	submission	13105 13687	87.1

Table 3: Predicted vs actual start/finish times based on average phone duration

Iterating through wordspotting and checking in order to bootstrap to an increasingly comprehensive speech database

As can be seen from Table 2, the next phase of the procedure involves using an annotation-assistance tool to refine the alignment over the first 1000 words which can then be used for training several variations

of two types of wordspotters, HMM-based wordspotters constructed using the HTK toolset [6] and statistically-based wordspotters [7]. These wordspotters and rule-based phone-event spotters are then run in parallel. Some automatic consistency checking of the various results is then done (in particular to eliminate likely false fires and to signal likely transcription errors) and then the full results are inspected with the annotation-assistance tool and another portion of the alignment is adjusted. This means that a longer portion of text and speech is now correctly aligned and this, in turn, can be used to construct still further-improved wordspotters which are then run over the next portion of the file. This procedure is repeated as often as necessary until all the speech and text for each speaker is aligned. As the wordspotters become more robust and comprehensive, the amount of manual adjustment decreases and, thus, the speed of database production increases.

Moving towards speaker independence

When enough speech recording and text for a large enough number of speakers is in the database, the material can be pooled to train speaker-independent wordspotting to help align recording and text for speakers for whom we have less than 5000 words per speaker. As yet we have not carried out this part of the overall procedure.

REFERENCES

- [1] John Butzberger, Hy Murveit, Elizabeth Shriberg, Patti Price, "Spontaneous Speech Effects in Large Vocabulary Speech Recognition Applications", DARPA Workshop on Speech and Natural Language Processing, February 1992, Morgan Kaufmann, pp. 339-343.
- [2] Sheryl Young and Michael Matessa, "Using Pragmatic and Semantic Knowledge to Correct Parsing of Spoken Language Utterances", Proc. Eurospeech 91, Genova, September 1991, pp. 223-227.
- [3] M.S.Schmidt and G.S.Watson, "The Evaluation and Optimization of Automatic Speech Segmentation", Proc. Eurospeech 91, Genova, September 1991, pp. 701-704.
- [4] S.Fujiwara, Y.Komori and M.Sugiyama, "An Integrated System for Automatic Labelling Based on HMM and Spectrogram Reading Knowledge", ISSPA 92, Signal Processing and its Applications, Gold Coast, August 1992, pp. 275-278.
- [5] Brit van Ooyen, Anne Cutler and Pier Marco Bertinetto, "Click Detection in Italian and English", Proc. Eurospeech 93, Berlin, September 1993, pp. 681-684.
- [6] S.J.Young, *HTK Version 1.4: Reference and User Manual*. Cambridge University Engineering Dept., Speech Group, August 1991.
- [7] M.O'Kane and P.Kenne, "Word and Phrase Spotting with Limited Training", Proc. Eurospeech 93, Berlin, September 1993, pp.1269-1272.