Automated Captioning of Television Programs: Development and Analysis of a Soundtrack Corpus

Ingrid Ahmer
Robin W. King

Institute for Telecommunications Research, University of South Australia
The Levels Campus, Mawson Lakes, SA 5095, Australia

ingrid@spri.levels.unisa.edu.au
robin.king@unisa.edu.au

Abstract

The purpose of this research is to investigate methods for applying speech recognition techniques to improve the productivity of off-line captioning for television. We posit that existing corpora for training continuous speech recognisers are unrepresentative of the acoustic conditions of television soundtracks. To evaluate the use of application specific models to this task we have developed a soundtrack corpus (representing a single genre of television programming) for acoustic analysis and a text corpus (from the same genre) for language modelling. These corpora are built from components of the manual captioning process. Captions were used to automatically segment and label the acoustic soundtrack data at sentence level, with manual post-processing to classify and verify the data. The text corpus was derived using automatic processing from approximately 1 million words of caption text.

The results confirm the acoustic profile of the task to be characteristically different to that of most other speech recognition tasks (with the soundtrack corpus being almost devoid of clean speech). The text corpus indicates that application specific language modelling will be effective for the chosen genre, although a lexicon providing complete lexical coverage is unattainable. There is a high correspondence between captions and soundtrack speech for the chosen genre, confirming that closed-captions can be a useful data source for generating labelled acoustic data. The corpora provide a high quality resource to support further research into automated speech recognition.

1. INTRODUCTION

The overall objective of this research is to devise signal-processing techniques to assist in the off-line production of television captions for hearing impaired viewers. Captioning is an international industry and the demand for captioned television programs continues to increase. In Australia, for example, recent legislation that requires all television programs to be captioned from the year 2001 will result in a threefold increase in the amount of captioning. As the demand for captioning increases, techniques for improving the productivity of the captioning process become more important, and so it is natural to start investigating automated techniques. The core task in television captioning is speech transcription, and the core operation for automated speech transcription is automatic speech recognition. We wish to consider the extent to which speech recognition technologies might improve the productivity of captioning.

Television captioning is a demanding application for automatic speech recognition (ASR). Fully automated captioning would require very high performance speaker-independent recognition of continuous natural speech with an unrestricted vocabulary and robustness against background music and environmental sounds. The difficulty of the task has been highlighted by recent experiments to quantify the performance of speech recognition on broadcast video and audio material. Wactlar et al. (1998) reported a word error rate of 50-65% for the dialog portions of documentary videos, increasing to 75% for a full hour documentary video. Although a lower word error rate of 20% was reported for clean speech recorded by a professional narrator in a TV studio, this performance is still quite unsatisfactory as a transcription error target for automated caption production [8].

In order to investigate the challenges in the transcription of television programs, representative labelled training data is required for experimentation and testing. Existing publicly available speech corpora for training continuous speech recognisers are not representative of the acoustic conditions of

general television programming. It is thus desirable to have additional sources of training and test data in order to assess the problems characterising a particular transcription task and to determine the value of application-specific models for improving acoustic segmentation and transcription performance.

Television captions contain an (almost verbatim) transcription of what was said on a program soundtrack and are time-aligned to the corresponding acoustic segments. Manual captioning of television programs for the benefit of the hearing impaired has been performed in Australia for over 20 years for a wide range of program types including news and current affairs, documentaries, dramas and quiz shows etc. Caption data would seem to offer a rich source of training data for ASR transcription covering several genres if methods could be developed for converting such data (and corresponding soundtracks) into useful corpora.

Some attempts have been made to use caption data to collect acoustic training data for recognisers. Witbrock & Hauptmann (1998) describe an unsupervised method for extracting accurately transcribed sections of broadcasts by matching closed-caption text to the outputs of a speech recogniser [9]. Only 4.5% of the spoken words were extracted by this method and these words were from sections of speech that could already be transcribed accurately by the recogniser. To investigate the problems for ASR transcription for television programming, we desired to have labelled acoustic data characterising the *entire* soundtrack.

Our initial goal was to develop a soundtrack corpus, representing a single genre of television programming, for acoustic analysis and to develop a text corpus from the same genre for language modelling. Subsequent sections of the paper describe the development and analysis of both a soundtrack corpus and a text corpus derived from the products of manual captioning. We provide comparisons of these corpora with existing speech and language databases to demonstrate the unique characteristics of the captioning task. Most notable is the very small proportion of clean speech samples occurring in the corpus. Nevertheless, the properties of the text corpus suggest that good results may be obtainable using application specific language models for this genre of text; however, a lexicon providing complete lexical coverage for the task may be unattainable.

2. CORPORA DEVELOPMENT

2.1 Source Data

Audio and text data were provided to the project by the Australian Caption Centre, and represented 10 hours from the category “travel programs”. This genre was selected for the research project as there are two such programs currently in production, guaranteeing availability of sufficient data for the project, and because the speaking rate for these programs rarely exceeds 3 words per second, thereby maximising the likelihood of verbatim caption text.

Video-cassette tapes of the program episodes had the program soundtrack recorded in mono on audio track 1 and a time-code in SMPTE format [6] recorded on audio track 2. Corresponding transcription files were in the European Broadcasting Union (EBU) data exchange format [4]. In addition, a further 180 EBU format program transcriptions (approximately 1 million words) were provided to be used for application specific language modelling.

2.2 Soundtrack Corpus

Corpus Preparation

To derive sentence level segmentation of the data, the soundtrack was first digitally sampled in stereo. The channels were separated and the time-code audio signal decoded, using software, to derive signal time offsets for each video frame number. Caption text with video frame numbers was extracted from the EBU files and used to annotate the audio signal, using the time offsets. To improve the quality of the data so that it would be suitable for speech research applications, the labelling was manually verified for transcription and time alignment accuracy and the acoustic characteristics of each segment were recorded.

Description of Soundtrack Corpus

The acoustic data is stored on CD-ROM in both ESPS and ESIG formats, and can be accessed as either an entire program soundtrack or as individual speech and non-speech segments. The annotation data is stored on disk in Entropic *waves+* label format. Sentence level labelling was automatically derived from the caption data and manually checked. Word and phonemic level labels are intended to be derived automatically using aligners. Cross-reference tables identify the parameters for each segment. In addition, a user interface has been developed to allow access to the annotated sound data by selecting the data characteristics of interest.

The data classification parameters characterise the acoustic dimensions of this corpus, but have also been designed to offer compatibility with other (less specific) general audio data category systems (refer to Tables 1 & 2 below). Both speech and non-speech signal data are identified to provide training data for acoustic segmentation techniques. The non-speech segments have been classified into six disjoint sound classes (1 silence, 2 music-vocal, 3 music–instrumental, 4 vocal-event, 5 noise and 6 other). Speech segments have been classified to identify eight parameters characterising acoustic conditions and speaker characteristics (1 whether speaker is a program presenter, 2 sex of speaker, 3 intensity of background music, 4 intensity of background noise, 5 accent, 6 spontaneity of speech, 7 sound quality, 8 speaker name).

2.3 Text Corpus

Corpus Preparation

All text normalisation was performed automatically using programs written to address the conventions of Australian subtitling [2] and the specific conditions encountered in the text. Text not representing spoken utterances was removed (eg. sound effects, speaker identifiers, song lyrics etc) and digits and symbols were converted to text. Sentence boundaries were identified and other punctuation symbols were removed. Hyphenated words were decompounded but apostrophes were retained within words (with the exception of "*'s*" which is treated as a separate word). Case distinction was retained to facilitate more accurate language modelling but this necessitated extra processing to decapitalise sentence initial words where the normal usage was lower-case.

Description of Text Corpus

The text corpus consists of 937,130 words representing 80,716 sentences of normalised text. The application specific lexicon derived from the text corpus contains 27,702 words. Capitalised words make up 7.88% of the corpus text but account for 32.67% of the lexicon. An analysis of the text corpus is provided in Section 3.2.

3. CORPORA ANALYSIS

3.1 Acoustic Characteristics of the Data

Detailed acoustic analysis has been performed on two program soundtracks at this stage and these results are reported here. Speech samples make up 69% of the acoustic data.

The most outstanding acoustic characteristic of the soundtrack corpus is the scarcity of clean speech. Only 2.1% of the speech samples represent studio quality clean speech. Most of the remainder have speech superimposed with other sounds, with 66.5% of the speech samples containing background music and 40.8% containing environmental sounds.

These results are in marked contrast to the corpora derived from broadcast news programs. Table 1 provides a comparison with the Hub-4 Radio Broadcast News data. The latter contains over 70% clean speech compared to our 3.7%. Spina & Zue (1996), in analysing general audio data derived from radio news, reported that clean speech accounted for 54% of the acoustic data in their corpus [7]. This compares with clean speech being only 2.5% of our acoustic data, as shown in Table 2.

Category of Speech Data	ARPA Hub 4	Soundtrack Corpus
F0: Baseline broadcast speech	33.73%	2.1%
F1: Spontaneous broadcast speech	16.03%	0.8%
F2: Speech over telephone channels	20.69%	0.0%
F3: Speech in the presence of background music	4.68%	50.4%
F4: speech in degraded acoustic conditions	9.84%	25.0%
F5: Speech from non-native speakers	0.76%	0.8%
FX: Speech not falling into other categories	14.27%	21.5%

Table 1: Speech classification comparison with ARPA Hub 4 data [3].

Category of Audio Data	GAD	Soundtrack Corpus
c_s: clean speech	54%	2.5%
f_s: field speech (bandwidth limited)	7%	0.0%
m_s: music speech	10%	34.8%
n_s: noise speech	12%	17.3%
m: music	6%	15.9%
sil: silence	10%	9.7%
gar: garbage	1%	19.8%

Table 2: Sound classification comparison with MIT Spoken Language Group GAD (general audio data) [7]. (Note that, in deriving the comparison, we have classified speech segments with *both* background music *and* background noise in the 'garbage' category.)

The speech samples in the soundtrack corpus, in general, represent clearly articulated prepared speech with grammatically complete utterances and very few disfluencies. They are, however, often delivered in an *expressive* manner. The majority of utterances were spoken by program presenters (95%), offering some scope for the development of speaker adaptive recognition techniques. Verification of the automatically produced sentence labels showed that 75% of the sentences had been transcribed verbatim in the captioning process, so that language models derived from large volumes of caption text (for this genre of program) would be well matched to the data.

3.2 Lexical Characteristics of the Data

The data in Table 3 show that the text corpus is characterised by high lexical coverage by relatively small sets of words. For general English text, a set of 141 words provides 50% lexical coverage [5]. However, for our particular genre of text, a set of two thirds this size (94 words) will suffice. Less than 4% of the lexical entries (1084 words) account for 80% of word usage. 11% of the lexical entries offer 90% corpus coverage, improving to 95% coverage when 25% of the lexical entries are used. To increase coverage past 98%, however, words would have to be used which have only occurred once or twice in the corpus.

Words occurring only once or twice in the corpus account for over 51% of the lexical entries and, of these, 38% are capitalised. Each hour of caption text contains approximately 60-80 distinct words not occurring in any other episode. This represents an expected out-of-vocabulary (OOV) rate of about 2% for each new program. Analysis of the infrequently occurring words shows that most of the lower case words are simply inflected forms of common dictionary words which should be expected to occur in a comprehensive lexicon for English speech. There are also a variety of phonetic spellings mimicking spontaneous speech and some foreign words. The capitalised words derive primarily from the names of places, people and businesses. Some were simply capitalised versions of common words (e.g. Disabled Motorists Association) but many truly represented obscure names of people and places (e.g. Tibrogargan, Undjaranjara).

Level of Corpus Coverage	Subset Size No. Words	Subset Size % of Distinct Words	Lowest Corpus Frequency for Words in Subset
50%	94	0.3%	1,333
60%	200	0.7%	606
70%	445	1.6%	255
80%	1,084	3.9%	91
90%	3,284	11.8%	22
95%	7,072	25.5%	7
98%	13,634	49.2%	2
100%	27,702	100.0%	1

Table 3: Lexical coverage for the text corpus of 937,130 words using subsets of the lexicon of 27,702 words.

For many applications, complete lexical coverage is unachievable, and proper nouns and spontaneous speech forms can present particular difficulties [1], [5]. In our case we expect to be able to reduce the OOV rate for common English words by supplementing the lexicon from other sources. Nevertheless it is unlikely that a lexicon can be developed which will completely overcome the OOV problem for many of the types of proper nouns that are typical of this task.

4. DISCUSSION

Utility of Caption Sources

Manual verification of captions has allowed detailed evaluation of caption sources. High transcription accuracy was found, including representation of spontaneous speech expressions and speech disfluencies. Non-verbatim captioning occurs primarily when the speaker rate exceeds three words a second (necessitating abridged caption text) and when appropriate text is already included in the video image. Captions must be matched to visual as well as speech events and this will cause some lack of synchronisation between captions and utterances.

We found that 75% of the utterances in the sound track have been transcribed verbatim in the caption text. 6% of the utterances had not been transcribed at all due to the data already occurring in the video image. The transcriptions of the remaining 19% required only small changes, many of which (eg. removal of sound effects) would not have been required if automatic text normalisation had been performed on this data prior to manual verification.

In their research, Witbrock & Hauptmann [9] rejected the direct use of closed-captions for acoustic segmentation and labelling due to a very high word-error-rate in their source captions (15.7% reported). Consequently, their methods were able to extract less than 5% of the available speech data from their sources. In our work, in contrast, high transcription accuracy has allowed us, with moderate manual post-processing, to utilise the entire program soundtrack and to create a well defined corpus of speech and other acoustic data characterising all aspects of the program genre.

Reliable unsupervised methods of extracting acoustic training data are usually preferred to manual approaches. Because of the high transcription accuracy of this task, we believe that the most productive approach to the automatic identification correctly transcribed speech segments, for this data, is likely to come from the use of automatic aligners to match caption text to acoustic segments. We are currently developing aligners trained on noisy data to provide word and phonemic segmentation for the corpus. We intend to test the potential for these aligners to automatically derive acoustic training data from the source captions and soundtracks.

Implications for Automatic Speech Recognition

The most important acoustic challenge with this type of data will be the development of effective noise reduction techniques and/or noise-tolerant ASR. In current work to develop aligners trained on noisy data, we are artificially generating phonetically labelled utterances which are acoustically similar to the corpus by adding characteristic sound effects from the non-speech segments to phonetically labelled

clean speech. This increases the range and quality of data available for training and testing noise-tolerant acoustic recognition models.

The text corpus should provide useful data for application specific language modelling, evidenced by the high lexical coverage achievable by relatively small sets of words. The application specific lexicon derived from the corpus is inadequate, however, and additional sources will be needed for deriving a sufficiently comprehensive lexicon to accommodate common English words encountered in broadcast material. Techniques for reliably recognising out-of-vocabulary words will be essential in ASR applications for this material as the introduction of new proper nouns (eg. obscure place names) is characteristic of this genre.

5. CONCLUSION

This initial research has demonstrated the feasibility of using existing closed-caption sources for generating labelled acoustic data suitable for speech recognition research. We have developed corresponding soundtrack and text corpora for one program genre. Although limitations imposed by the captioning process mean that verbatim transcription is not guaranteed in the caption text (and so some verification is required) the captions for our chosen genre proved to be a valuable data resource readily convertable into a format suitable for speech processing. Our corpora have the potential to be a high quality resource to support further research in automated caption transcription.

6. ACKNOWLEDGMENTS

This work is being conducted with the support of the Australian Caption Centre (ACC) through the ARC SPIRT grant scheme. The authors acknowledge, in particular, the ACC's provision of caption and soundtrack material and their general advice on captioning processes

7. REFERENCES

1. Adda-Decker, M. & Lamel, L. 1997, 'The use of lexica in automatic speech recognition', *Proceedings ELSNET's 5th International Summer School on Language and Speech Communication*, Leuven, Belgium.

2. Australian Caption Centre 1997, *Manual of Subtitling Standards*, Supertext Offline Edition, ACC, Sydney.

3. Cook G.D., Kershaw D.J., Christie J.D.M., Seymour C.W. & Waterhouse S.R. 1997, 'Transcription of Broadcast Television and Radio News: the 1996 Abbot System', *Proceedings IEEE-ICASSP-97*, Munich, pp 723-6.

4. European Broadcasting Union 1991, *Specification of the EBU subtitling data exchange format*, Tech.3264-E, Geneva.

5. Liberman, M.Y. & Church, K.W. 1991, 'Text Analysis and Word Pronunciation in Text-to-Speech Sythesis', in *Advances in Speech Processing*, ed. Furui, S. & Sondhi, M.M., Marcel Dekker Inc, New York.

6. SMPTE 1995, *SMPTE Standard for Television, Audio and Film – Time and Control Code*, no. ANSI/SMPTE 12M-1995, White Plains, NY

7. Spina M.S & Zue V.W. 1996, 'Automatic Transcription of General Audio Data: Preliminary Analyses', *Proceedings ICSLP-1996*, Philadelphia, pp 594-7.

8. Wactlar H. D., Hauptmann A.G. & Witbrock M.J.1998, *Infomedia News-On-Demand: Using Speech Recognition to Create a Digital Video Library*, Technical Report CMU-CS-98-109, Carnegie Mellon University, Pittsburgh.

9. Witbrock M.J. & Hauptmann A.G. 1998, *Improving Acoustic Models by Watching Television*, Technical Report CMU-CS-98-110, Carnegie Mellon University, Pittsburgh.