GENERATING AND EVALUATING SEGMENTATIONS FOR AUTOMATIC SPEECH RECOGNITION OF CONVERSATIONAL TELEPHONE SPEECH

S. E. Tranter, K. Yu, G. Evermann, P. C. Woodland

Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, UK.
Email: {sej28,ky219,ge204,pcw}@eng.cam.ac.uk

ABSTRACT

Speech recognition systems for conversational telephone speech require the audio data to be automatically divided into regions of speech and non-speech. The quality of this audio segmentation affects the recognition accuracy. This paper describes several approaches to segmentation and compares the resulting recogniser performance. It is shown that using Gaussian Mixture Models outperforms an energy-detection method and using the output from the speech recogniser itself increases performance further. An upper bound on possible performance was obtained when deriving a segmentation from a forced alignment of the reference words and this outperformed using manually marked word times. Finally the correlation between an appropriately defined segmentation score and WER is shown to be over 0.95 across three data sets, suggesting that segmentations can be evaluated directly without the need for full decoding runs.

1. INTRODUCTION

Automatic Speech Recognition (ASR) systems for conversational telephone speech (CTS) data require the audio to be divided into regions of speech and non-speech. The quality of the audio segmentation affects the performance of the recogniser, for example if speech regions are labelled as non-speech and discarded they produce deletion errors, whereas regions of non-speech which are not discarded in segmentation *may* produce insertion errors. Recognition performance may also be affected by other properties of the segmentation such as the minimum/maximum duration of a speech segment or the tightness of the segment boundaries, for example by affecting the normalisation or adaptation.

This paper investigates several different methods of producing segmentations, based on either acoustic information, such as an adaptive energy-based method or using Gaussian Mixture Models (GMMs); or on word-level timing information, such as using a recogniser output, a forced alignment of the reference words, or manually generated word times. The segmentations are compared both from the resulting recognition word error rates (WER) and a segmentation score which is the sum of the missed speech and false alarm rates compared to a reference segmentation. The results are given from experiments into how to maximise the correlation between the segmentation score and WER, so as to allow the WER to be predicted solely from the segmentation without the need for potentially computationally expensive decoding runs.

This paper is arranged as follows. Section 2 explains the 'diarisation' score used to evaluate the segmentations, Section 3 discusses the data used in the experiments and Section 4 describes the segmentations. Section 5 briefly describes the recognition systems used to generate the WERs and compares the WERs resulting from the segmentations, then Section 6 evaluates the correlation between the diarisation score and the WER. Finally conclusions are offered in Section 7.

2. THE DIARISATION SCORE

The segmentations are evaluated using the diarisation score, which was defined for the 2003 Spring Rich Transcription (RT-03s) diarisation evaluation [1]. The general formulation takes a reference and a hypothesis segmentation and performs a one-to-one mapping of the reference speaker IDs to the hypothesis speaker IDs so as to maximise the total overlap of the reference and (corresponding) mapped hypothesis speakers. Speaker detection performance is then expressed in terms of the miss (speaker in reference but not in hypothesis), false alarm (speaker in hypothesis but not in reference), and speaker-error (mapped reference speaker is not the same as the hypothesised speaker) rates. The overall diarisation score is the sum of these three components, and can be calculated using the following formula:

$$\frac{\sum_s dur(s) \cdot (\max(N_R(s), N_H(s)) - N_C(s))}{\sum_s dur(s) \cdot N_R(s)}$$

where s is the longest continuous piece of audio for which the reference and hypothesised speakers do not change, $dur(s)$ is the duration of s, $N_R(s)$ is the number of reference speakers in s, $N_H(s)$ is the number of hypothesised speakers in s and $N_C(s)$ is the number of mapped reference speakers which match the hypothesised speakers. For the CTS data, the channels are provided separately with only one speaker per side, so this becomes:

$$\frac{\sum_s dur(s) \cdot (H_{miss}(s) + H_{fa}(s))}{\sum_s dur(s) \cdot H_{ref}(s)}$$

where H is always zero except $H_{miss}(s)$ is 1 for a missed speech segment, $H_{fa}(s)$ is 1 for a false alarm speech segment and $H_{ref}(s)$ is 1 for a segment containing a reference speaker. Thus missed speech and false alarm speech errors are weighted *equally* in the error count.

In the RT-03s diarisation evaluation, regions which contained speaker-attributable vocal noise (and surrounding silence) were excluded from scoring. We do not do this since, for the purposes of ASR, any events in the audio which are not speech should be treated in the same way as silence. We also chose to generate our reference segmentations in a slightly different way (see section 6 for more details).

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

In *Proc. ICASSP 2004*, Volume I, pp. 753–756.

3. DATA USED IN EXPERIMENTS

The experiments reported in this paper were conducted on the development and evaluation data used in the English CTS RT-02 and RT-03 Rich Transcription evaluations.[2] The main data sets are the RT-02 evaluation data *(eval02)* consisting of 60 5-minute conversations, the December 2002 RT-03 dryrun data set *(dry03)* which is a 12 conversation subset of eval02, and the main RT-03 STT evaluation set *(eval03)* consisting of 72 5-minute conversations. The eval02 and eval03 sets are relatively large and provide more reliable results to be obtained, but the dry03 subset is also necessary since some transcriptions or segmentations were only available on this data. In particular, manually derived word times, which were initially used to derive the diarisation reference were only produced on the dry03 data set. Further details about the exact composition of the data sets can be found in [3].

4. GENERATING SEGMENTATIONS

The data must be segmented into speech and non-speech regions for recognition. The CUED RT-03 CTS recognition system used a segmentation based on the CUED RT-03s CTS diarisation system. This used GMMs to segment and label the audio as silence, male or female and ran in 0.05xRT. Two different model sets were built, one using Switchboard-I and II (phase 1 and 2) data, and the other using cellular (Switchboard-II phase 4) data. Approximately 3 hours of training data was used for each model and the silence model contained 128 mixture components, whilst the male and female models both contained 256. A Viterbi decoder was used to find the most likely sequence of GMMs, ensuring only a single gender and dataset per side were postulated. An insertion penalty was used to prevent rapid oscillation between models. Each side was processed independently, so no cross-channel modelling such as that used by BBN [4] was performed, although experiments suggest that the standard of the CUED and BBN CTS RT-03 segmentations is similar. Further details are given in [3].

Additional segmentations were generated in the following ways

CUED Pre-ASR: This includes the GMM based system described above, and a number of similar systems with slight variations in training data, models and/or parameters.

CUED Post-ASR: The word times output by the speech recogniser are used to define the regions of speech. The primary run, `Post-ASR-full`, used the CUED RT-03 187xRT CTS recogniser [5] to generate the word times. A contrast run, `Post-ASR-fast` using the CUED RT-02-based 10xRT recogniser described in section 5 to generate the word times is also provided.

Baselines: These are the two baseline segmentations, rt02base and rt03base, provided by MIT-LL for the Rich Transcription evaluations [2]. They use an adaptive energy-based detector [6] and differ considerably in quality.

CUED FA: A forced alignment of the reference words to the audio is performed and the resulting word times used to define the regions of speech.

Manual word times: For the Dec 2002 dryrun 12-side subset of the eval02 data, manually produced word times were provided, which were used to define the regions of speech. Non-lexical tokens were ignored.

These initial segmentations were refined by having segments of silence which were less than a certain critical length relabelled as speech (smoothing) and where applicable, the boundaries of speech segments expanded (padding). The smoothing and padding parameters for diarisation scoring were chosen so as to match those used when generating the reference file. It was found empirically that using 0.6s smoothing and 0.2s padding resulted in the lowest WER on the dry03 data, giving a 7.2% relative gain over the case of no smoothing or padding [3]. Therefore when the segmentations are used as input to the speech recogniser, 0.6s smoothing and 0.2s padding are added unless otherwise stated.

5. GENERATING WORD ERROR RATES

The recognition system used in this paper to measure the effect of different segmentations on recognition accuracy is based on the 2002 CUED 10xRT CTS system developed for the RT-02 STT evaluation [7]. The acoustic models were improved but the system structure was unchanged. The system uses cross-word triphone models and a fourgram language model with a 54k dictionary. The acoustic features were based on PLP analysis and normalised using VTLN.

The system operates in three passes. The initial pass uses relatively simple models to generate a transcription for use as supervision in the estimation of VTLN warp factors and global MLLR adaptation transforms. The following two passes use models trained using Minimum Phone Error Estimation and employ a global HLDA transform in the feature extraction. The second pass generates lattices which are then rescored in the final stage using models adapted using 2 MLLR speech transforms per speaker.

In an additional experiment the more sophisticated 2003 CUED 10xRT system [8] was used on the eval03 data. This system employs two separate acoustic model sets in separate branches of the final rescoring stage whose outputs are combined. One model set was trained using Speaker Adaptive Training and the other employs a single pronunciation dictionary (SPron). All models were trained using MPE and used HLDA. Adaptation was performed using lattice-based MLLR and full-variance transforms.

The WERs for the segmentations described in section 4 on the eval02 and dry03 subset, and the eval03 data are given in Table 1 along with that from using the (manually defined) segmentation used in ASR scoring (the STM file) with no *additional* smoothing or padding. These results show that the GMM-based pre-ASR segmentation consistently outperforms the energy-based baseline segmentations and the post-ASR segmentation outperforms the pre-ASR segmentation on the eval02 and eval03 data sets, showing that segmentations can be improved using the ASR output. (Since the dry03 subset data set is only a fifth of the eval02 data set, the WER numbers are more reliable on the latter).

It is also interesting to note that the segmentation derived from the CUED forced alignment times consistently provide the best WER results, outperforming the segmentation derived from the manually marked word times on the dry03 subset, and the manually defined STM segmentation on all three data sets. This may be down to a system interaction effect, but suggests that to get an upper bound on segmentation performance, or to predict the WER of a CUED recognition system from just comparing segmentations, the reference segmentation should be derived from the CUED forced alignment times.

System	dry03	eval02	eval03	
BASELINE rt02base	29.5	29.2	27.1	(22.8)
BASELINE rt03base	28.3	28.0	26.7	(22.4)
CUED Pre-ASR	28.1	27.3	26.3	(22.2)
CUED Post-ASR-fast	28.0	27.2	26.2	(22.0)
CUED Post-ASR-full	28.2	27.1	26.0	(22.0)
CUED FA word times	27.4	26.2	25.4	(21.3)
Manual word times	27.8	—	—	(—)
STM (unknown smth/pad)	27.7	26.7	25.6	(21.6)

Table 1. Word Error Rates using different segmentations for CTS data. Numbers in parenthesis are from using the RT-03 based recogniser.

6. THE CORRELATION BETWEEN DIARISATION SCORE AND WER

The quality of the segmentation clearly affects the quality of the recogniser output and thus the WER. Ideally we would like to be able to predict the WER directly from the segmentation, so as to allow different segmentation configurations to be tried without needing to perform a (computationally expensive) full decode for every case. The diarisation score offers a way of measuring segmentation performance by summing the missed speech and false alarm speech giving equal weighting to both. This does not therefore reflect the commonly held view that for ASR segmentations the missed speech is more important than the false alarm speech, since the latter is recoverable for example by matching a silence acoustic model. However, it offers an unbiased comparison of two segmentations in that the numerator of the error score is independent of which file is the reference and which the hypothesis.

In order to investigate the correlation between the diarisation score and the WER, 16 segmentations were made on the dry03 data subset. These were derived from 10 CUED Pre-ASR runs, 2 CUED Post-ASR runs, 2 Baseline runs, the CUED forced alignment word times and the manual word times. A diarisation reference was generated from the manual word times using 0.6s smoothing, and diarisation scores of the (similarly smoothed) segmentations were calculated. Segmentations were similarly made on the eval02 data for all cases except for the manual word times (which were only available on the dry03 subset). The WER was then found using the RT-02 based recogniser described in section 5 on the dry03 data and the eval02 superset after adding an additional 0.2s padding to the (smoothed) segmentations. The results are illustrated in Figure 1.

It was noted in section 5 that the CUED-FA derived segmentation gave a lower WER than using the manual times possibly due to a system interaction effect, and thus it may be more appropriate to use the CUED-FA times to derive the diarisation reference when trying to use the diarisation score to predict the WER of a CUED recogniser. To investigate this more carefully, the diarisation scores were recalculated using a reference derived from the CUED forced alignment word times. The results are illustrated in Figure 2. The correlation coefficients between the diarisation scores and the word error rates are given in Table 2.

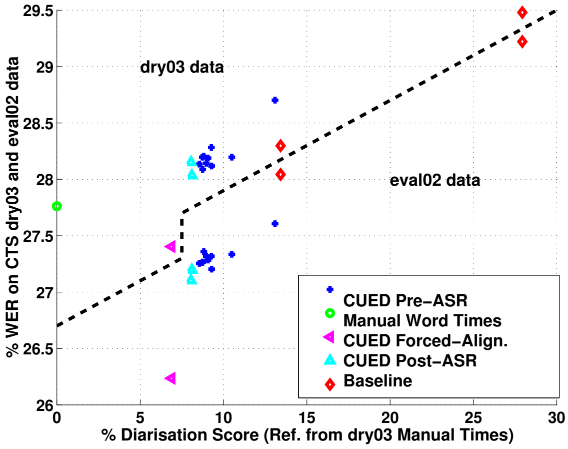

Fig. 1. Relationship between diarisation score on dry03 data and WER on eval02 and dry03 data. The dashed line shows the division between the two data sets. The manual word times were used to derive the diarisation reference.

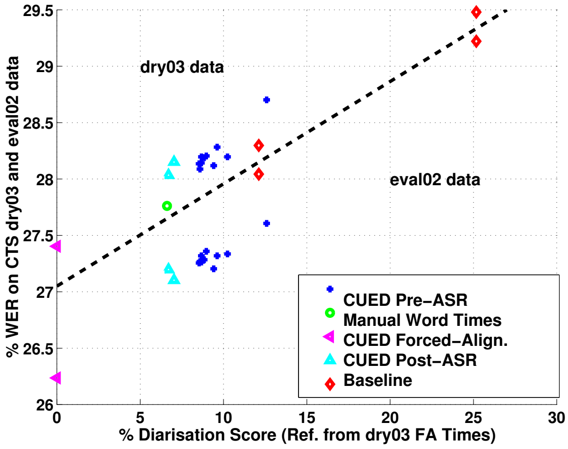

Fig. 2. Relationship between diarisation score on dry03 data and WER on eval02 and dry03 data. The dashed line shows the division between the two data sets. The CUED forced alignment word times for the dry03 data were used to derive the diarisation reference.

	DIARY Manual Ref	DIARY FA Ref	WER dry03	WER eval02
DIARY(Man)	1.00	0.94	0.90	0.93
DIARY(FA)	-	1.00	0.98	0.98
WER(dry03)	-	-	1.00	0.94
WER(eval02)	-	-	-	1.00

Table 2. Correlation Coefficients for predicting the eval02 WER from the dry03 subset

The results show there is a strong correlation between the diarisation scores and the subsequent WERs of the system, and this correlation is highest when the diarisation reference is derived from the CUED forced alignment word times. In particular, the correlation between the diarisation score and the WER on the dry03 data rises from 0.90 to 0.98, and that between the dry03 data diarisation score and the WER on the eval02 superset rises from 0.93 to 0.98. This is very encouraging, given that the correlation between the WERs themselves on the two sets is only 0.94. This suggests that predicting the WER on the eval02 superset of data using just the dry03 subset, can be done with as much confidence using the diarisation score as the WER, so new segmentations can be tested without the need for computationally expensive decoding runs.

An added advantage of using forced alignments to generate the diarisation reference is that they can be produced with relatively little manual effort and thus, unlike using manually generated word times, are possible to obtain for large data sets. A diarisation reference was thus constructed for the eval02 and eval03 data starting with the CUED forced alignments. The results are illustrated in Figure 3 and the correlation between diarisation scores and WER given in Table 3. These results confirm that there is a very high correlation between the diarisation score and the WER providing the reference is generated appropriately, and this correlation is maintained across different data sets and recognisers.

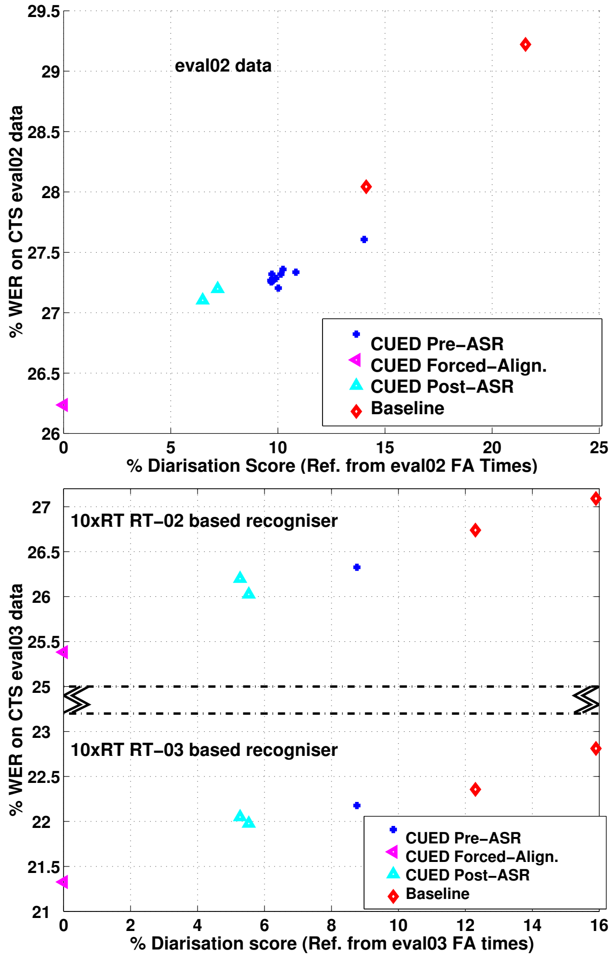

Fig. 3. Relationship between diarisation score and WER on the eval02 and eval03 data, using the CUED forced alignment word times to derive the diarisation reference.

Dataset	Size	Num Points	Correlation
eval02	5 hrs	15	0.96 / 0.96
eval03	6 hrs	6	0.99 / 0.99 (0.97 / 0.98)

Table 3. Correlation Coefficients between the diarisation score and the WER when using the CUED forced alignment to derive the diarisation reference. The second number is when adding 0.2s padding (in addition to the 0.6s smoothing) to the segmentations when calculating the diarisation score. The numbers in parenthesis are from using the RT-03 based recogniser.

7. CONCLUSIONS

Segmentations for the CTS data can be generated using many different methods and can be compared using the diarisation score. Adding 0.6s smoothing and 0.2s padding to the segmentations minimised the resulting WER from the recogniser and thus 0.6s smoothing was added to the segmentations for diarisation scoring. Also including the padding was found to have little impact on the correlation between the diarisation scores and WER and so was omitted.

Using the ASR output to refine the segmentation proved beneficial, reducing both diarisation score and WER. An upper bound on performance was obtained using a segmentation derived from the CUED forced alignment word times. This also outperformed using manually derived word or segment-level times. Generating the diarisation references from the CUED forced alignment word times rather than the manually derived word times also increased the correlation between diarisation score and WER, and made it possible to score much larger data sets.

The WER on the eval02 data can be predicted from the dry03 subset with just as much confidence using the diarisation score as the WER itself. The correlation coefficient between the diarisation score and WER was over 0.95 on both the eval02 and eval03 data sets even when the recogniser was changed, showing the value of a segmentation for ASR can generally be judged from the diarisation score without needing computationally expensive recogniser runs.

8. REFERENCES

[1] NIST, "The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan, version 4," http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt03-spring-eval-plan-v4.pdf, 25th February 2003.

[2] NIST, "Benchmark Tests : Rich Transcription (RT)," http://www.nist.gov/speech/tests/rt/.

[3] S.E. Tranter, K. Yu, D.A. Reynolds, G. Evermann, D.Y. Kim, and P.C. Woodland, "An Investigation into the Interactions between Speaker Diarisation Systems and Automatic Speech Transcription," Tech. Rep. CUED/F-INFENG/TR-464, Cambridge University Engineering Department, October 2003.

[4] D. Liu and F. Kubala, "A Cross-Channel Modeling Approach for Automatic Segmentation of Conversational Telephone Speech," in *Proc. ASRU*, St. Thomas, U.S. Virgin Islands, December 2003, pp. 333–338.

[5] G. Evermann, H. Y. Chan, M. J. F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P. C. Woodland, "Development of the 2003 CU-HTK Conversational Telephone Speech Transcription System," in *Proc. ICASSP*, 2004.

[6] D. A. Reynolds, P. Torres, and R. Roy, "EARS RT03s Diarization," in *Proc. Spring 2003 Rich Transcription Workshop (RT-03s)*, Boston, MA, May 2003.

[7] T. Hain, P. C. Woodland, G. Evermann, X. Liu, G. L. Moore, D. Povey, and L. Wang, "Automatic Transcription of Conversational Telephone Speech - Development of the CU-HTK 2002 System," Tech. Rep. CUED/F-INFENG/TR-465, Cambridge University Engineering Department, December 2003.

[8] G. Evermann and P. C. Woodland, "Design of Fast LVCSR Systems," in *Proc. ASRU*, St. Thomas, U.S. Virgin Islands, December 2003, pp. 7–12.