The segmentation of multi-channel meeting recordings for automatic speech recognition

John Dines [a] Jithendra Vepa [a]
Thomas Hain [b]

IDIAP–RR 06-22

April 2006

published in
Interspeech-ICSLP 2006, Pittsburgh

[a] IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL), Martigny, Switzerland
[b] Department of Computer Science, University of Sheffield, UK

The segmentation of multi-channel meeting recordings for automatic speech recognition

John Dines Jithendra Vepa Thomas Hain

April 2006

published in
Interspeech-ICSLP 2006, Pittsburgh

Abstract. One major research challenge in the domain of the analysis of meeting room data is the automatic transcription of what is spoken during meetings, a task which has gained considerable attention within the ASR research community through the NIST rich transcription evaluations conducted over the last three years. One of the major difficulties in carrying out automatic speech recognition (ASR) on this data is dealing with the challenging recording environment, which has instigated the development of novel audio pre-processing approaches. In this paper we present a system for the automatic segmentation of multiple-channel individual headset microphone (IHM) meeting recordings for automatic speech recognition. The system relies on an MLP classifier trained from several meeting room corpora to identify speech/non-speech segments of the recordings. We give a detailed analysis of the segmentation performance for a number of system configurations, with our best system achieving ASR performance on automatically generated segments within 1.3% (3.7% relative) of a manual segmentation of the data.

1 Introduction

The segmentation of recordings is an important preprocessing step for automatic speech recognition (ASR) in that it provides a means to divide the signal into manageable portions for the decoder, provides a basis for the reliable application of channel normalisation schemes such as cepstral mean and variance normalisation (CMN, CVN), and also contributes to reduce the errors by restricting processing to only those portions of the recording that should be transcribed. In the domain of multi-channel meeting recordings from independent headset microphones (IHM) this is a particularly important task, since on average each channel will only have speech activity inversely proportional to the number of meeting participants. Moreover, during non-speech segments there is often a high-level of 'noise' on each participant's microphone, due mostly to either cross-talk from adjacent speakers or non-speech sounds (such as breath and laughter), which leads to a much higher word error rate (WER) than that produced using a segmentation which removes all irrelevant portions of the recording.

The sustained interest in automatic transcription of meeting recordings has seen a considerable amount of research devoted to the automatic segmentation of IHM recordings [1, 2, 3]. Drawing on a variety of approaches, this work has demonstrated that segmentation of these recordings using traditional voice activity detection (VAD) approaches is generally insufficient and extra care is particularly needed to deal with the presence of cross-talk. While demonstrating improvements over simple VAD schemes, this work has still exhibited substantial performance gaps of 8% to 10% relative WER in comparison to ASR carried out on the reference segmentation, motivating further investigation of this problem.

In this paper we present the speech segmentation system for multi-channel meeting recordings which was developed as part of the AMI meeting transcription system [4]. This system was used in the AMI consortium's submission to the the NIST rich transcription 2006 Spring meeting recognition evaluation[1]. The focus of this work has been not only to minimise the impact of automatic segmentation on WER, but to thoroughly investigate our approach in order to better appreciate the relationship between segmentation and ASR performance. Towards this end we have conducted a number of experiments that explore this in detail.

The remainder of the paper is organised as follows. In Section 2 we give a brief overview of the meeting room resources used in this work. Section 3 describes our proposed segmentation system, including feature extraction, training of the speech/non-speech classifier and the segmentation of speech data. Section 4 describes experiments that were conducted on the NIST rich transcription spring 2005 meeting room task (RT05s), including a detailed discussion of the results. Finally, in Section 5 we give some concluding remarks and future directions of this work.

2 Meeting room resources

2.1 Training resources

Training data for the ASR models and speech segmentation system is sourced from a number of corpora; namely, the ICSI [5], NIST [6], ISL [7] and AMI [8] corpora which have been collected over a number of years within the framework of different projects. These corpora total approximately 112 hours of speech over 150 meetings. For the training of the segmentation only 8 minutes of each meeting was used with a proportion of speech to non-speech frames of approximately 1:5.5.

2.2 Evaluation resources

Over the last three years (2004-06) NIST has conducted rich transcription evaluations focused on the meeting room domain. The Augmented Multi-party Interaction (AMI) consortium submitted a system for the RT05s evaluation [4], which comprises 120 minutes of meeting data from six sites including

[1] see http://www.nist.gov/speech/tests/rt/rt2006/spring/

those in the training sets as well data recorded at Virginia Tech [9]. In this paper we use experiments conducted on the RT05s meeting room evaluation data with the 2005 AMI meeting recogniser's first pass models and the Juicer large vocabulary speech decoder [10]. Due to the present limitations of our WFST decoder, a pruned language model was used. Further details of the AMI meeting transcription system and Juicer decoder may be found in the referenced papers.

3 Proposed system

In our formulation of the speech segmentation problem, we consider that speech segmentation output should closely replicate the conditions under which the ASR models were trained. In order to achieve this it is possible to use the ASR models themselves to determine the speech/non-speech segments of the training data, then use this as the basis for training the speech activity detection classifier. We note that this approach is similar in spirit to that in [11], except that where their work uses the actual ASR acoustic models in the speech activity detection, we prefer to train a separate classifier for this task, enabling us to incorporate additional auxiliary features not used during acoustic model training to deal specifically with cross-talk.

3.1 Features

The features used in the ASR system, 12 MF-PLP features and c_0, plus their first and second derivatives, and an additional energy feature (giving 42 dimension vector) were used as the baseline for our speech segmentation classifier [12]. In addition, we wished to experiment with an additional set of auxiliary features to aid in the detection of cross-talk. Cross-talk is notoriously difficult to discriminate from the actual target speech using information derived from a single channel alone, hence, we introduce a cross-meeting normalised energy feature which incorporates the energy of all N IHM channels as follows:

$$E_i^{norm}(n) = \frac{E_i(n)}{\sum_{k=1}^{N} E_k(n)} \tag{1}$$

where $E_i(n)$ is the signal energy for channel i at frame n. Thus, it can be seen that $E_i^{norm}(n)$ measures the energy on the current channel relative to the energy across all channels, independently of the actual recording level since it is bounded between zero and one.

We use three additional features, signal kurtosis, mean cross-correlation and maximum normalised cross-correlation, which have been previously demonstrated to be useful for cross-talk detection [3]. Further details of these features can be found in the referenced article. With the addition of these auxiliary features and their first and second derivatives, an input feature vector of dimension 54 is generated.

3.2 Classifier training

The segmentation system is built around a multi-layer perceptron (MLP), with hyperbolic tangent hidden activation function and softmax output activation function, trained on speech/non-speech target classes. Based on a number of preliminary experiments, an input layer of 31 consecutive input frames and 50 hidden units was found to give a good balance of accuracy and complexity.

Training targets are generated by performing a forced alignment of the training data using the ASR models and reference segments. All labels from the forced alignment excepting the 'sil' model are labelled as the speech target class and the 'sil' model label and inter-segment parts of the meeting recording are labelled as the non-speech target class. Thus, using the ASR models and reference segments, we have designated the training data as speech/non-speech in a manner which is consistent with the ASR acoustic models.

Training of the MLP proceeds by taking the meetings and splitting these into training and validation sets with a ratio of 9:1. The features from the training set are normalised to have zero mean and unit variance and the transformation which generates this normalisation is retained for application to the validation data and later to the test data. Training of the MLP parameters is carried out using standard error back propagation of the Kullback-Leibler divergence criterion with early stopping being determined by the validation set.

3.3 Cross-talk suppression

During testing, the MF-PLP features are calculated in two ways. One set of features us calculated using the original IHM meeting recording and the second set is calculated on IHM recordings which have been preprocessed by a cross-talk suppression algorithm. This cross-talk suppression is based upon a modified version of the adaptive-LMS echo cancellation [13]. Modifications were carried out to enable the use of multiple reference channels and also to account for the difference in recording conditions between the meeting and telephone channels for which the algorithms was originally developed. The auxiliary features were only calculated on the original recordings since they are explicitly intended to aid in the detection of cross-talk.

3.4 Segmentation

Segmentation of test data is carried out using Viterbi search in which scaled likelihoods are generated from the MLP classifier class posterior estimates and class prior probabilities:

$$p(x(n)|\mathcal{C}_j) = \frac{P(\mathcal{C}_j|x(n))}{P(\mathcal{C}_j)} \tag{2}$$

where $x(n)$ is the vector of concatenated speech features and $\mathcal{C}_j, j \in \{sp, sil\}$ represents the speech and non-speech classes, respectively. Segment minimum duration imposed via the HMM topology of the speech/non-speech models. Segment minimum duration, M, log insertion penalty, I, and class priors can all be tuned on development data to optimise the system for performance at the frame level or WER. In our system we use a fixed $M = 50$ and $I = -15$, though in practice the system does not appear to be overly sensitive to these settings. Lastly, smoothing of the segments is carried out by adding a 100 ms collar to the output of the Viterbi search, with merging of any segments with less than 200 ms gap.

4 Experiments

4.1 Experimental setup

Automatic segmentation is evaluated by comparing reference and automatic segments at the frame level and also by comparing subsequent ASR performance using the reference and automatically generated segments. A set of reference segments for the RT05s evaluation are provided by NIST, which have been manually transcribed and, as such, do not necessarily correlate well with what would be obtained via automatic means. To test this we performed forced alignment of the evaluation data using the reference transcripts and generated a segmentation of the data based upon the forced-alignment. In comparing these segments at the frame level we observe that 4% of frames are in discrepancy, however, the WER obtained using the original reference segments and the segments derived from forced alignment are 35.5% and 35.2%, respectively. We postulate that this improvement comes from the calculation of CMN and CVN statistics that are more consistent the training of our ASR system and demonstrates that there is the (albeit slim) potential for automatic segmentation to exceed that of manual transcribers. In the following frame level comparisons we use the forced-alignment segments as the reference segmentation.

Segmentation systems were evaluated using two MLP classifiers; one trained from MF-PLP features alone and a second that also incorporates the auxiliary features (+AUX). We note that the second

MLP comprises approximately 23% more parameters, though, we shall endeavour to show that the differences between these systems is largely derived from the features rather than the number of parameters. Thus, we obtain a total of four distinct systems by generating MF-PLP features calculated with (+CT) and without the cross-talk suppression previously described:

System A MF-PLP

System B MF-PLP+CT

System C MF-PLP+AUX

System D MF-PLP+CT+AUX

4.2 Results

In the first set of tests we calculated the percentage of frames falsely recognised as speech (FA) and falsely recognised as silence (FR) relative to the total number of frames for each of the four systems against the reference segmentation for different speech/non-speech class prior probabilities. The results of these tests are shown in Figure 1.

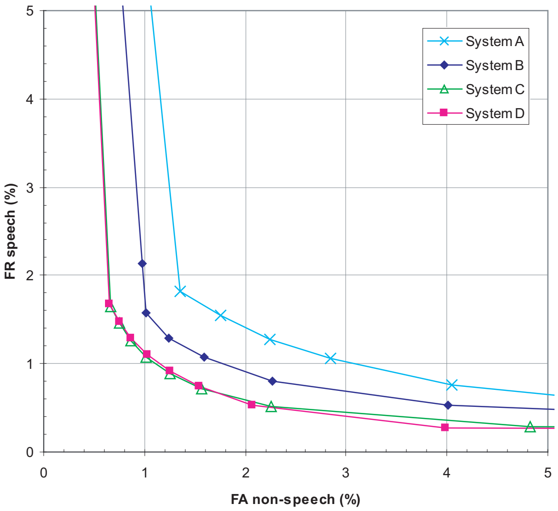

Figure 1: Frame level performance of auto-segmentation systems against reference segments generated from forced alignment.

We next ran our recogniser on segments generated by systems B and D for different speech/non-speech class prior probabilities. The results of these tests are shown in Table 1 and Figure 2.

4.3 Analysis

The direct optimisation of the segmentation system using the ASR output would be desirable, but is practically infeasible except for the tuning of a few hyper-parameters. Hence, in our system the bulk of parameters are optimised using a frame level criterion. The relationship between frame level accuracy and WER is still of primary importance, since, if we can demonstrate a direct relationship between the two, then we can postulate that optimisation at the frame level will lead to better performance with respect to WER. This favours the MLP classification approach which we have adopted, since, the training criterion attempts to minimise frame classification error rate. In the following discussion

System	$P(C_{sp})$	FA	FR	Sub	Del	Ins	WER
REF$_m$	—	—	—	23.1	8.9	3.5	35.5
REF$_{fa}$	—	0	0	22.2	10.2	2.9	35.2
B	0.40	1.0	2.1	20.7	15.0	3.1	38.8
	0.25	1.2	1.3	21.8	12.2	3.8	37.8
	0.15	2.3	0.8	22.4	10.6	5.7	38.7
D	0.40	0.7	1.7	21.0	13.6	3.1	37.7
	0.25	**1.0**	**1.1**	**21.5**	**11.7**	**3.6**	**36.8**
	0.15	1.5	0.7	22.3	10.5	4.4	37.2

Table 1: ASR and frame-level performance for different speech/non-speech segmentations. REF$_m$ refers to the manual segmentation and REF$_{fa}$ refers to the segmentation produced using forced alignment.

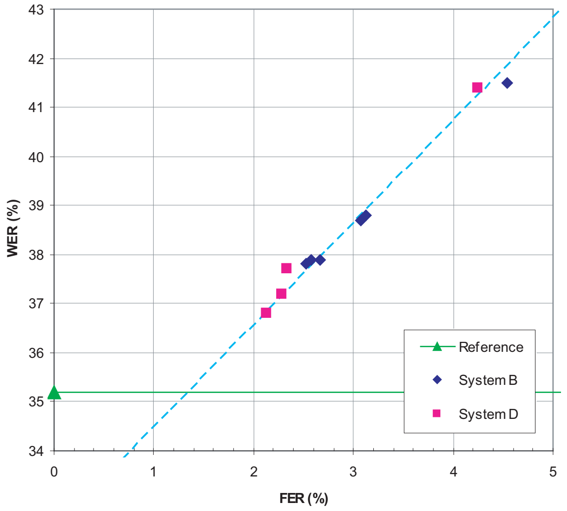

Figure 2: Word error rate (WER) versus frame error rate (FER = FA + FR) for automatically generated segments.

we explore this relationship and attempt to show that this is indeed the case for the approach we have presented.

Several points are immediately apparent from an examination of Figure 1. First of all we note that the cross-talk suppression has a significant impact on reducing the false acceptance rate when using MF-PLP features alone (Systems A and B), whereas with the addition of the auxiliary features the cross-talk suppression has almost no impact (Systems C and D). This observation supports our hypothesis that the addition of auxiliary features and not the increase in number of parameters is the main factor in distinguishing the two MLPs that were trained. It is also clear that the use of auxiliary features has a greater impact on performance than the cross-talk suppression.

We also note that the minimum frame error rate occurs at the equal error rate of each of the systems, suggesting that this should be the optimal operating point - from a frame error perspective - which may be chosen using development data. Next, looking at Table 1 and Figure 2 we see that there is indeed a linear relationship between frame error rate and word error rate and that false acceptance and false rejection errors have equal impact on WER, though, these errors manifest differently in the form of insertion and deletion errors, respectively, which may be of some importance depending on how the ASR output is used.

Lastly, we highlight that the best WER performance was achieved using System D with $P(C_{sp}) = 0.25$. This system achieves a WER degradation or 1.3% (3.7% relative) and 1.7% (4.8% relative) against the manual and forced-alignment reference segments, respectively.

5 Conclusions

We have presented a system for the automatic segmentation of IHM multiple channel meeting recordings. The system is based upon an MLP classifier, with the best results obtained using traditional ASR features combined with auxiliary features to aid in the detection of cross-talk. Evaluating this system using state-of-the-art ASR we achieved a 1.3% (3.7%) degradation in WER in comparison to a manually derived segmentation of the RT05s meeting room evaluation data. We have also presented detailed results and discussion which provide useful insight into the segmentation performance with respect to a reference segmentation at both the frame level and WER from an ASR system.

In future work we will continue to develop our system along several lines. Firstly, further investigation of appropriate auxiliary features is called for, along with further refinements to our training regime. We also intend to investigate the joint decoding of all IHM channels for a meeting simultaneously as proposed in [3]. Such an approach is well suited to our segmentation paradigm as posterior probabilities from individual channels may be simply combined to give joint-state probabilities.

6 Acknowledgements

The authors would like to thank all those involved in the development of the AMI ASR system, which formed the basis of the evaluations. This work was supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811) and the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2.

References

[1] T. Pfau, D. Ellis, and A. Stolke, "Multispeaker speech activity detection for the ICSI meeting recorder," *Proc. ASRU*, December 2001.

[2] K. Laskowski, Q. Jin, and T. Schultz, "Crosscorrelation-based multispeaker speech activity detection," in *Proc. ICSLP*, Jeju Island, Korea, 2004.

[3] S. Wrigley, G. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, January 2005.

[4] T. Hain, L. Burget, J. Dines, I. McCowan, M. Karafiat, M. Lincoln, D. Moore, G. Garau, V. Wan, R. Ordelman, and S. Renals, "The development of the AMI system for the transcription of speech in meetings," in *Proc. MLMI*, Edinburgh, UK, 2005.

[5] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. ICASSP*, Hong Kong, 2003.

[6] J. S. Garofolo, C. D. Laprun, M. Michel, V.M. Stanford, and E. Tabassi, "The NIST meeting room pilot corpus," in *Proc. 4th Intl. Conf. on Language Resources and Evaluation (LRECŠ04)*, 2004.

[7] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *Proc. ICSLP*, 2002.

[8] J. Carletta, S. Ashby, S. Bourban, M. Guillemot, M. Kronenthal, G. Lathoud, M. Lincoln, I. McCowan, T. Hain, W. Kraaij, W. Post, J. Kadlec, P. Wellner, M. Flynn, and D. Reidsma, "The AMI meeting corpus," in *Proc. MLMI*, 2005.

[9] "Spring 2005 (RT-05S) rich transcription meeting recognition evaluation plan," http://www.nist.gov/speech/tests/rt/rt2005/spring/rt05s-meeting-eval-plan-V1.pdf, May 2005.

[10] D. Moore, J. Dines, M. Magimai Doss, J. Vepa, O. Cheng, and T. Hain, "Juicer: A weighted finite state transducer speech decoder," in *Proc. MLMI (to appear)*, Washington DC, May 2006.

[11] E. Marcheret, K. Visweswariah, and G. Potamianos, "Speech activity detection fusing acoustic phonetic and energy features," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 241–244.

[12] P. C. Woodland, M. J. F. Gales, D. Pye, and S. J. Young, "Broadcast news transcription using HTK," in *Proc. ICASSP*, Munich, Germany, 1997, pp. 719–722.

[13] D. Messerschmitt, D. Hedberg, C. Cole, A. Haoui, and P. Winship, "Digital voice echo canceller with a TMS32020," Application report SPRA129, Texas Instruments, 1989.