A semi-Markov model for speech segmentation with an utterance-break prior

Mark Sinclair[1]*, Peter Bell*[1]*, Alexandra Birch*[2]*, Fergus McInnes*[1]

[1]Centre for Speech Technology Research, [2]Statistical Machine Translation Group
School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK
{mark.sinclair, a.birch, peter.bell, fergus.mcinnes}@ed.ac.uk

Abstract

Speech segmentation is the problem of finding the end points of a speech utterance for passing to an automatic speech recognition (ASR) system. The quality of this segmentation can have a large impact on the accuracy of the ASR system; in this paper we demonstrate that it can have an even larger impact on downstream natural language processing tasks – in this case, machine translation. We develop a novel semi-Markov model which allows the segmentation of audio streams into speech utterances which are optimised for the desired distribution of sentence lengths for the target domain. We compare this with existing state-of-the-art methods and show that it is able to achieve not only improved ASR performance, but also to yield significant benefits to a speech translation task.

Index Terms: speech activity detection, speech segmentation, machine translation, speech recognition

1. Introduction

We define speech segmentation as the problem of finding the end points of a speech utterance in time. While this may at first seem like a relatively simple goal it is in fact a non-trivial problem to define. As speech does not strictly follow the same rules we find in written language, such as sentence breaks, it can often be highly subjective as to what constitutes an appropriate segmentation of speech for a given task. The vagueness and high-order decision processes that surround these concepts make it challenging to design an effective automatic speech segmentation system.

Most automatic speech segmentation methods work by identifying speech and non-speech regions based on acoustic evidence alone e.g. contrasting energy levels or spectral behaviour[1] [2] [3]. Some more recent research has improved upon this foundation by using richer feature sets that are more suited to the task or include long-term dependances [4] [5] [6]. Others have begun to apply deep learning techniques which can garner more discriminative features and improve robustness [7] [8] [9]. However, all of these methods still only consider the acoustics and are not necessarily exploiting the underlying structure of the spoken language.

Human transcribers, on the other hand, are capable of segmenting speech by exploiting a greater wealth of prior information such as syntax, semantics and prosody in addition to such acoustic evidence. As a result, human transcribers may opt to ignore acoustically motivated 'breaks' in speech in favour of maintaining longer segments based on semantic knowledge. Such an informed segmentation can greatly influence subsequent system components that have been explicitly designed to exploit the patterns and structure of natural language, e.g. the language models used in automatic speech recognition (ASR) or machine translation (MT).

Previous work on detecting segmentation of sentence like units has looked at using features such as prosody [10], language model scores [11] [12], translation model scores [13] and syntactic constituents [14]. [15] presents a review of some of this work and also motivates tuning the segmentation of speech to the task at hand as we do in this paper. Our approach of modelling global sentence length distribution is orthogonal to much of this previous work, and combining these information sources would be beneficial. There has been some previous work which attempts to exploit some of these cues [16]. However, in the present paper, we have limited our focus to the use of statistics of utterance durations and present a novel way of exploiting this to select a globally optimal sequence from acoustically motivated'break candidates'. We find that this yields an advantage over the use of local acoustic information alone at putative pauses in the speech. We also find indications that the optimal setting of the segmentation parameters varies with the ultimate task (e.g. transcription or translation) that is to be achieved using the segmented speech. We present results on segmentation, recognition and translation of TED talks[1].

2. Utterance-break Modelling

While the automatic speech segmenters that we initially used are only able to segment on an acoustic basis, they would actually perform this task very well. When compared to the manual segmentation we found the False Alarm rate to be very low (2-3%) while the more dominant error is Missed Speech.

Empirical evidence suggests that the automatic segmenters work very well at framewise classification but are not able to distinguish when a non-speech segment is simply a pause inside a speaker's utterance or a true 'break' between utterances as judged by human annotators. Often such pauses are quite short and as such a naïve approach might be to simply alter the minimum duration constraint for non-speech regions. However, in practice we find that this simply shifts the balance from Missed Speech to False Alarm errors by removing more potential breaks, quickly resulting in over-long segments. A significant part of this behaviour is due to the fact that such systems are only able to make local decisions about whether or not to include a non-speech segment. To remedy this problem, we propose to investigate methods for globally optimising the sequence of utterance breaks, incorporating prior knowledge about the likelihood that non-speech breaks should be included given their temporal relationship i.e. the duration between them.

2.1. Break Candidates

As a precursory step to find the globally optimal sequence of utterance breaks B^*, we first need to derive a sequence of can-

[1]http://www.ted.com

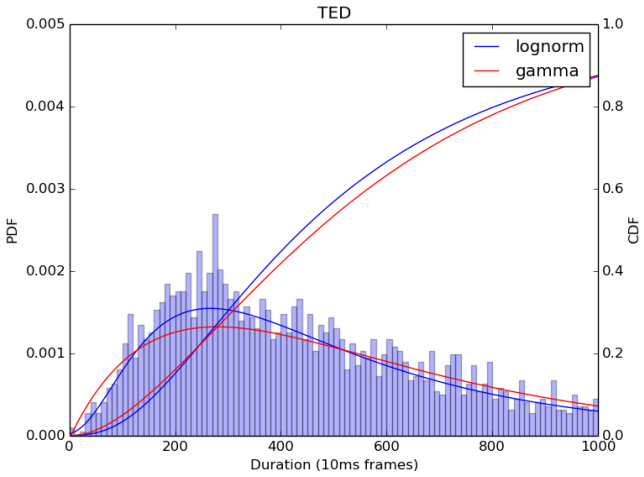

Figure 1: *Log-normal vs. Gamma PDFs fitted to speech segment durations of TED dev set. The CDFs provide the prior likelihood of a new break given the duration after the last break.*

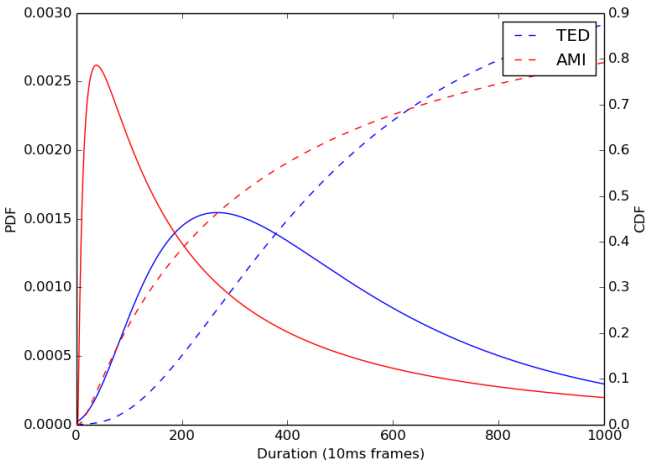

Figure 2: *Log-normal PDFs and corresponding CDFs for speech segment durations of TED vs AMI data.*

didate break points B. Ideally the candidate sequence should be broad enough that it includes a good optimal sequence as a sub-set so we would therefore require that $|B| \geq |B^*|$, where $|B|$ and $|B^*|$ are the cardinalities of the candidate and optimal sequences respectively. The candidates themselves can be determined by any kind of initial segmentation method such as an existing acoustically motivated speech segmentation algorithm.

2.2. Utterance-break Prior

In order to make decisions about whether or not to include a candidate break, we need to know the prior probability of a break, which we condition on the time since the last break was observed. This may be derived from a statistical model of segment durations. Figure 1 shows a histogram of speech segment durations for a development set of lecture data. We investigated the use log-normal and gamma distributions to represent the behaviour of the data and ultimately chose the former as it provided a slightly better fit as shown in Figure 1. To derive the break likelihood prior we simply use the cumulative density function (CDF) of this distribution as shown in Equation 1 where d is the duration since the previous break and α is a scaling factor to account for the difference in dynamic range compared to the acoustic likelihood.

$$f_{brk\_utt}(d) = (\frac{1}{2} + \frac{1}{2}\text{erf}\left[\frac{\ln d - \mu}{\sqrt{2}\sigma}\right])^{\alpha} \quad (1)$$

We were interested in how such a prior may vary according to the domain. Therefore, as a contrast to the prepared, rehearsed, single-speaker speech that is found in TED talks, we also looked at the distribution of speech segments for a set of AMI scenario meetings. These meetings contain multiple speakers discussing a given common task whereby the speech is unprepared and spontaneous. From Figure 2 we can observe that the distribution has a much lower mean, illustrating the intuition that speech segments are generally shorter during such dynamic group discourse. This suggests that the break likelihood prior could be adapted for different domains to achieve optimal performance.

2.3. Viterbi Decoding

In order to determine the globally optimal sequence of utterance breaks B^* we consider a sequence of candidate breaks $B = \{b_0, \ldots, b_{|B|}\}$ as a semi-Markov process whereby each candidate is a state. We can then perform Viterbi decoding over a sparse $|B| \times |B|$ trellis, whereby each position i moves, not through uniform time segments, but through the indices of B. Each position j allows us to consider the transition arriving at break b_i from b_j. As the break candidate states are only forward connected we can only arrive at a given break from a previous break, hence $j < i$. We keep a vector of tuples T that records the start and end frame indices of each break, this allows us to calculate the duration between any pair of break candidates $d_{i,j} = t_{i,start} - t_{j,end}$.

We also use the sums of the frame-level log-likelihoods from the speech/non-speech segmenter for acoustic features X, where x_i is a vector representing all the frames that are in break b_i. We use this to create a posterior probability P_{brk_aco}, as shown in Equation 2, that represents the acoustic probability of a given break. The purpose of the normalisation is that we want the acoustic likelihood of breaks to increase with duration so that long breaks are favoured.

$$P_{brk\_aco}(i) = \frac{\ell_{nonspch}(x_i)}{\ell_{nonspch}(x_i) + \ell_{spch}(x_i)} \quad (2)$$

Therefore, the probability of the partial sequence that has a break at i is formalised as

$$v_i = \max_{j<i} [v_j + \log f_{brk\_utt}(d_{i,j})] + \log P_{brk\_aco}(i) \quad (3)$$

with $v_0 = 0$. For each candidate i we store the identity of the state j which maximises Equation 3. This allows to B^* to be recovered by a backtrace procedure.

As the duration between the break candidate under consideration and earlier ones increases, it will become very unlikely that such long term transitions would occur. In practice, we can therefore afford to prune the lattice by ignoring transitions from earlier states that are further away than a prescribed maximum segment duration, δ. As such, for each index i we only consider transitions from states where $d_{i,j} \leq \delta$. This is illustrated in Figure 3.

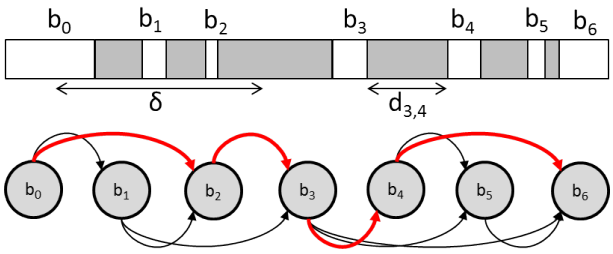

Figure 3: *An example of a candidate break sequence and associated state topology. We can see that the states can only feed forward and some long-term transitions have been pruned such as $b_0 \rightarrow b_3$ as $d_{0,3} > \delta$. The transitions highlighted in red show an example optimal break sequence $B^* = \{b_0, b_2, b_3, b_4, b_6\}$*

3. Experiments

3.1. Data

For evaluation, we used the data made available for the IWSLT evaluation campaigns[17]. This comprises a series of TED talks that have been divided into development sets (dev2010 and dev2011) and a test set (tst2010), each containing 8-11 talks. The talks average just under 10mins in length and each contains a single English speaker (either native or non-native). The talks are manually segmented and transcribed at the utterance level. We also had manual English-French translations for evaluating the MT system component.

3.2. Speech Segmentation Systems

3.2.1. Manual

Here we simply pass the manual segmentation to the ASR and MT systems directly in order to form the oracle standard with which to compare our automatic speech segmenters.

3.2.2. SHOUT

This system makes use of the SHOUT Toolkit (v0.3)[2][18] which is a widely-used off-the-shelf speech segmentation system. The tool uses a GMM-HMM-based Viterbi decoder, with an iterative sequence of parameter re-estimation and re-segmenting. Minimum speech and silence duration constraints are enforced by the number of emitting states in the respective HMMs.

3.2.3. Baseline segmenter

Our baseline system, labelled "simple" in the tables, is identical to that used for our recent lecture transcription system [19] and comprises a GMM-HMM based model which is used to perform a Viterbi decoding of the audio. Speech and non-speech are modelled with diagonal-covariance GMMs with 12 and 5 mixture components respectively. We allow more mixture components for speech to cover its greater variability. Features are calculated every 10ms from a 30ms analysis window and have a dimensionality of 14 (13 PLPs and energy). Models were trained on 70 hours of scenario meetings data from the AMI corpus using the provided manual segmentations as a reference. A heuristically optimised minimum duration constraint of 500ms is enforced by inserting a series of 50 states per class that each have a transition weight of 1.0 to the next, the final state has a self transition weight of 0.9.

[2]http://shout-toolkit.sourceforge.net/download.html

3.2.4. Break Smooth

Here we introduce our utterance-break prior model. In order to establish the candidate break sequence we use the system in Section 3.2.3 to do an initial segmentation pass over the data. The only exception is that the minimum duration constraint is reduced to 100ms. If used directly, this would normally perform very poorly as a VAD but when used as input to the subsequent break smoothing we have three advantages over the original constraint: better guarantee of enough candidates to find an optimal solution, the ability to find shorter speech segments ($\geq$100ms), and more accurate end-points for segments between 100-500ms. The break likelihood prior was trained on the speech segment durations of the dev2010 and dev2011 sets. The maximum segment duration δ is set to 30 seconds.

We have also shown results for 2 different operating points of the scaling factor α, 30 and 80. While this parameter is designed to mitigate for the difference in dynamic range with the acoustic model, we found it subsequently functioned as a form of segment duration tuning whereby a greater α results in more break smoothing and hence longer segments.

3.2.5. Uniform

As well as our automatic methods we also considered segmenting each talk into uniform speech segments of length N seconds, which is equivalent to having a break of zero length at every interval. This allowed us to check whether or not the benefit of our utterance-break prior may simply be due to an 'averaging' of the break distribution. As the ASR system is still able to do decoder-based segmentation within each given segment, this is also a way of measuring its influence. Here, longer uniform segments leave more responsibility to the decoder for segmentation and at $N = 300$, the maximum segment length for the ASR system, we effectively allow the decoder to do all the segmentation (with potentially a small error at the initial segment boundaries).

3.3. Downstream System Descriptions

3.3.1. Automatic Speech Recognition (ASR)

ASR was performed using a system based on that described in [19]. Briefly, this comprises deep neural network acoustic models used in a tandem configuration, incorporating out-of-domain features. Models were speaker-adapted using CMLLR transforms. An initial decoding pass was performed using a 3-gram language model, with final lattices rescored with a 4-gram language model.

3.3.2. Machine Translation (MT)

We trained an English-French phrase-based machine translation model using the Moses [20] toolkit. The model is described in detail in our 2013 IWSLT shared task paper [21]. It is the official spoken language translation system for the English-French track. It uses large parallel corpora (80.1M English words and 103.5M French words), which have been filtered for the TED talks domain. The tuning and filtering used the IWSLT dev2010 set.

The goal of our machine translation experiments is to test the effect that ASR segmentation has on the performance of a downstream natural language processing task. The difficulty

with allowing arbitrary segmentations in MT is that automatic evaluation is performed matching MT output with gold reference sentences which have their own manual segmentation. In order to evaluate translations which have different segmentations, we need to align the MT output segmentation with the reference. We use a tool provided by the Travatar [22] toolkit which aligns files with different segmentations. It searches for the optimal alignment according to the BLEU score. We use it to align our MT output with a variety of different segmentation models, to our gold reference with manual segmentations. We align each TED talk in the test set separately to maximize performance.

3.4. Gold Transcription Mapping

Any improvement a given segmentation provides to the ASR system could subsequently improve the performance of the MT system. However, this makes it difficult to infer how much of the MT performance gain is simply a consequence of a better source transcript as compared with the direct influence of the segmentation itself. To control for this, we used the ASR system to make a forced alignment of the manual transcription in order to gain word-level timing information. We were then able to map any of our given segmentations with the same gold-standard transcription.

4. Results

We present results for all of our end-to-end automatic systems in Table 1. Firstly, we note that our *Simple* segmenter is able to significantly outperform *SHOUT*, confirming that we have competitive acoustic segmenter with which to form the foundation of our experiments. We can then see that our *Break Smooth* segmenter is further able to improve the performance of both ASR and MT over the *Simple* segmenter.

The performance of the *Uniform 300s* system showed a strong performance for ASR, falling only slightly short of our best performing *Break Smooth* system. We attribute this to the nature of TED talks whereby there is typically very little non-speech (illustrated by the 6.49% FA of *Uniform 300s*), which itself mostly comprises periods of silence of small duration, which is implicitly segmented by the silence HMMs used by the decoder itself. In contrast, however, when we use a uniform segmentation for MT, we find that it does not perform as well despite the good ASR performance. As the MT system ideally expects sentence-like segments, a uniform segmentation will not be practical for these purposes. This also shows that a segmentation that works well for ASR may not necessarily work well for MT and vice-versa.

In order to fully control for the dependence MT has for the WER of the ASR transcript it receives, we have shown the results for when we map each of our segmentations to the force-aligned gold transcription in Table 1. First of all, from the variation in performance we can infer that segmentation does indeed have a direct effect on MT performance. However, in these conditions we find that the MT system favours the break smoothing algorithm with shorter segments than MT. Figure 4 shows how the prior and posterior distributions compare. We can see that when $\alpha = 30$ the distribution takes a closer shape to the true distribution with a 'shift' to shorter segments, which could be due to the fact that the automated methods have more accurate segment boundaries. As such the MT system in this case could be benefitting from more 'sentence-like' utterances, whereas the ASR system can actually afford to have, and may actually benefit from, slightly longer segments as it is able to further segment in more detail using its own decoder.

	SAD		ASR	MT:ASR	MT:Gold
Segmentation	Miss	FA	WER	BLEU	BLEU
Manual	-	-	13.6	0.2472	0.2472
SHOUT	12.71	0.16	18.3	0.1967	0.2256
Simple	9.91	2.66	16.7	0.2007	0.2319
Break Smooth 30	4.25	2.33	15.0	0.2085	0.2409
Break Smooth 80	7.86	1.46	14.6	0.2104	0.2368
Uniform 300s	0.00	6.49	14.8	0.2014	0.2369

Table 1: Segmentation, ASR and MT results for each segmenter. MT results are shown for both ASR output and gold transcripts segmented with different segmentation models.

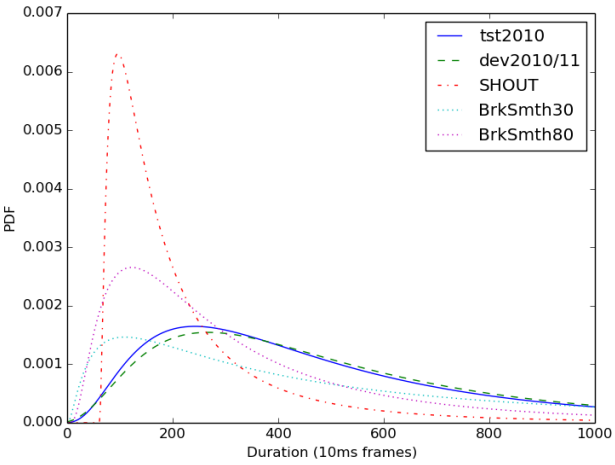

Figure 4: *A comparison of the prior and posterior segment length distributions.*

5. Conclusions and Future Work

We have shown that speech segmentation can be improved by exploiting non-acoustic prior knowledge – in this case, the use of an utterance-break model. Such improvements can be shown to propagate to further benefit the performance of downstream tasks such as ASR and MT. We have also shown that the benefits to MT are not simply a consequence of the benefits to ASR suggesting that speech translation performance is highly dependent on the quality of the speech segmentation. However, we observed that the optimal segmentations for each task are not necessarily the same – furthermore, typical speech segmentation evaluation metrics are not a reliable indicator of downstream system performance.

Given what we have learned from this investigation we believe there is scope in future work to add linguistic knowledge into the segmentation model, such as language modelling scores and even syntactic bracketing information. This would require running segmentation as an iterative procedure, on the output of an ASR model, before feeding it back in as the input to an ASR system.

6. References

[1] I. Boyd and D. K. Freeman, "Voice activity detection," Jan. 4 1994, uS Patent 5,276,765.

[2] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.

[3] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 139, no. 4, pp. 377–380, 1992.

[4] J. Ye, T. Kobayashi, M. Murakawa, and T. Higuchi, "Incremental acoustic subspace learning for voice activity detection using harmonicity-based features." in *INTERSPEECH*, 2013, pp. 695–699.

[5] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H. Hansen, A. Janin, B. S. Lee, Y. Lei, V. Mitra *et al.*, "All for one: feature combination for highly channel-degraded speech activity detection." in *INTERSPEECH*, 2013, pp. 709–713.

[6] A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, and S. Narayanan, "Multi-band long-term signal variability features for robust voice activity detection." in *INTERSPEECH*, 2013, pp. 718–722.

[7] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks." in *INTERSPEECH*, 2013, pp. 728–731.

[8] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 697–710, 2013. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6362186

[9] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 483–487. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6637694

[10] A. Stolcke, E. Shriberg, R. A. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words." in *ICSLP*, 1998.

[11] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 1005–1008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=607773

[12] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation." in *IWSLT*, 2006, pp. 158–165.

[13] E. Matusov, D. Hillard, M. Magimai-Doss, D. Z. Hakkani-Tür, M. Ostendorf, and H. Ney, "Improving speech translation with automatic boundary prediction." in *INTERSPEECH*, vol. 7, 2007, pp. 2449–2452.

[14] B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya *et al.*, "Reranking for sentence boundary detection in conversational speech," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1660078

[15] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tur, M. Harper, D. Hillard, J. Hirschberg, H. Ji, J. G. Kahn, Y. Liu *et al.*, "Speech segmentation and spoken document processing," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 59–69, 2008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4490202

[16] D. Rybach, C. Gollan, R. Schluter, and H. Ney, "Audio segmentation for speech recognition using segment features," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4197–4200.

[17] M. F. M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," *Proceedings IWSLT 2012*, 2012.

[18] M. A. H. Huijbregts, "Segmentation, diarization and speech transcription: surprise data unraveled," 2008.

[19] P. Bell, F. McInnes, S. R. Gangireddy, M. Sinclair, A. Birch, and S. Renals, "The UEDIN english ASR system for the IWSLT 2013 evaluation," *IWSLT, Heidelberg, Germany*, 2013.

[20] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the Association for Computational Linguistics Companion Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 177–180. [Online]. Available: http://www.aclweb.org/anthology/P/P07/P07-2045

[21] A. Birch, N. Durrani, and P. Koehn, "Edinburgh SLT and MT System Description for the IWSLT 2013 Evaluation," in *Proceedings of the 10th International Workshop on Spoken Language Translation*, Heidelberg, Germany, December 2013, pp. 40–48.

[22] G. Neubig, "Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers," in *Proceedings of the ACL Demonstration Track*, Sofia, Bulgaria, August 2013.