Automated speech-unit delimitation in spoken learner English

Russell Moore[1] **Andrew Caines**[1] **Calbert Graham**[1] **Paula Buttery**[2]
Automated Language Teaching & Assessment Institute
[1]Department of Theoretical & Applied Linguistics
[2]Computer Laboratory
University of Cambridge, Cambridge, U.K.
rjm49,apc38,crg29,pjb48@cam.ac.uk

Abstract

In order to apply computational linguistic analyses and pass information to downstream applications, transcriptions of speech obtained via automatic speech recognition (ASR) need to be divided into smaller meaningful units, in a task we refer to as 'speech-unit (SU) delimitation'. We closely recreate the automatic delimitation system described by Lee and Glass (2012), 'Sentence detection using multiple annotations', *Proceedings of INTERSPEECH*, which combines a prosodic model, language model and speech-unit length model in log-linear fashion. Since state-of-the-art natural language processing (NLP) tools have been developed to deal with written text and its characteristic sentence-like units, SU delimitation helps bridge the gap between ASR and NLP, by normalising spoken data into a more canonical format. Previous work has focused on native speaker recordings; we test the system of Lee and Glass (2012) on non-native speaker (or 'learner') data, achieving performance above the state-of-the-art. We also consider alternative evaluation metrics which move away from the idea of a single 'truth' in SU delimitation, and frame this work in the context of downstream NLP applications.

1 Introduction

By convention, texts written using the Latin alphabet are normally subdivided into smaller units – *sentences* – by capitalised initial characters and closing full-stops (periods), question marks and exclamation marks. The sentence has in turn become an orthodox unit of analysis for much linguistic research, from natural language processing to syntactic theory. Speech, meanwhile, is hardly ever so neatly portioned up. Pauses and turn-taking (in conversation) may at first appear to correspond to sentence-marking orthographic devices, and often they do delimit sensible language chunks, but not always. Speakers pause in strange places, make false starts, leave thoughts unfinished, and interrupt or overlap each other (Sacks et al., 1974; Dingemanse and Floyd, 2014; Carter and McCarthy, in press). These characteristic features of spontaneous speech pose a problem for researchers investigating monologues or dialogues in naturalistic scenarios: what is a sentence-like unit of spoken language?

This question has been addressed by conversation analysts, acquisition researchers assessing performance accuracy, and compilers of large corpora, among others. Common practice is to transform speech recordings into written transcripts in order to work further with the data, whether with manual or automated means. Now comes the dilemma of how to subdivide those transcripts, where appropriate, into chunks one can work with. For conversational data the first division is made by speakers' turns into what are known as 'utterances'. But then how should those utterances be further subdivided (if at all) into something akin to sentences?

The consensus, theoretically-speaking, has been to identify smaller, sentence-like units of speech on syntactic and/or semantic grounds (with more emphasis on the former), often with reference to prosody – *e.g.* intonation contour and pause duration. In terms of hand-annotated corpora, annotators are expected to have a 'feel' of where unit boundaries should go. For automated processing of large speech corpora,

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Page numbers and proceedings footer are added by the organisers. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 782–793, Osaka, Japan, December 11-17 2016.

researchers have attempted to model this human intuition using supervised machine learning algorithms (Shriberg et al., 2000; Roark et al., 2006; Favre et al., 2008).

In engineering research, this is one task in 'metadata extraction' (MDE), along with the detection of disfluencies and orthographic conventions related to the readibility of transcripts (Walker et al., 2005). However, to the best of our knowledge, no open source software was released with these studies. Moreover, these systems have been trained and tested on corpora of telephone conversations and news broadcasts produced by native speakers of English. Our data are monologues and so we seek to replicate the set-up reported by Lee and Glass (2012), who combined multiple information sources to delimit boundaries in monologue restaurant reviews.

Our contribution is firstly one of terminology, asserting that we are searching for SPEECH-UNITS – with 'speech-unit delimitation' (SU delimitation) our preferred name for this task, thereby avoiding reference to the 'sentence'. Note that this is not a new term, but it is not always made clear what the unit of speech analysis is, or what is meant by 'speech-unit' where it is used. Secondly, we release an SU delimitation toolkit in a public repository[1]. Thirdly, we consider alternative evaluation metrics for SU delimitation, to move away slightly from the idea of a single 'ground-truth'. Finally, we report how well our system performs on monologues produced by native speakers and learners of English, achieving an F-measure of 0.674 with our best performing set-up.

2 The speech-unit

The sentence-like unit of speech has a varied history in the applied linguistic field. Foster et al. (2000) offer a thorough review of past proposals, identifying three types: 'mainly semantic', 'mainly intonational', and 'mainly syntactic' units. They settle on the last type in their 'analysis of speech unit' (AS-unit), allowing multiple clauses in one unit, based on evidence from pause studies that syntactically-coherent units play a central role in speech planning (Deschamps, 1980; Raupach, 1980). We do not dispute these findings, but note that this unit is again heavily reliant on syntax for its definition, though it is held to reflect a psychological reality.

Meanwhile, the *de facto* standard analysis unit in conversation analysis (CA) is the 'turn construction unit', types of which are identified as "sentential, clausal, phrasal, and lexical – i.e. syntactically" (Sacks et al., 1974). Reference is made to 'sound production' and the ways it can disambiguate, for example, statements and questions. But otherwise, for CA the object of analysis is the transcript, and as such, with speech in written form, syntax is king.

Researchers at the LDC adopted a deliberately less precisely specified approach in adding punctuation to transcripts of speech: their chosen unit, the 'SU', "functions to express a complete thought or idea on the part of a speaker" (Strassel, 2003). What 'SU' actually stands for is left open: "some possibilities include: Sentential Units, Syntactic Units, Semantic Units and Slash Units" (Strassel, 2003). The SU is defined on the basis of both syntax *and* semantics so we will not prioritise either *a priori*. Finally, the 'slash unit' refers to a transcription convention that may be obscure to some, and we avoid such overt esotericism. Instead we think of an SU as a 'speech unit': a generic and sufficiently transparent concept. We also note that it has been used before by Roberts and Kirsner (2000) in their study of 'temporal cycles' in speech, defining it as, "a segmented part of speech and the hesitation pause that preceded it".

Again, we see allusion to the planning process here, except in this case semantics are brought to the foreground. LDC annotators were instructed to "rely primarily on semantic and syntactic information, and secondarily on prosodic information" when listening to the speech recordings and deciding where to delimit SUs. We see the benefit of this flexible approach, drawing on multiple information sources rather than mainly syntax, and we adopt the SU as our sub-unit of choice for speech. Once we have a reliable system to automatically identify SU boundaries it is of benefit to downstream tasks of two broad types: natural language processing (NLP) and computer-assisted language learning (CALL). For NLP tasks we want to know whether the SUs are sensible and expected in some way, and for CALL we require chunks of language which are useful for automated learner assessment and feedback.

[1] `http://github.com/rjm49/multistage_segmenter`

3 Speech-unit delimitation

Experiments in SU delimitation date back to work by Shriberg, Stolcke and colleagues (Stolcke and Shriberg, 1996; Stolcke et al., 1998; Shriberg et al., 2000). They initially developed a framework to bring together lexical, discourse and prosodic cues to tag transcripts with various kinds of hidden structural information, with some of these cues at first being hand-coded. They demonstrated that a combination of prosodic and language model features produced better tagging outputs than either feature type in isolation. Later, Shriberg et al. (2000) introduced fully automated extraction of the necessary cues. Since then, the systems have been extended with syntactic features to supplement *n*-gram models (Roark et al., 2006), alternative classifiers to the early decision tree (DT) models, such as the conditional random field (CRF) (Favre et al., 2008) and deep neural network (DNN) (Xu et al., 2014), ensemble models with multiple voting (Liu et al., 2005), and to languages other than English including Czech (Kolář et al., 2006) and Chinese (Tomalin et al., 2007). Xu et al. (2014) report leading results of 0.81 F-measure (harmonic mean of precision and recall) on SU boundaries with their DNN-CRF model, outperforming a DT-CRF baseline with 0.774 F-measure.

However, most of the above-mentioned systems have been trained and evaluated on native speaker telephone conversation and broadcast news dialogues: the Switchboard and Broadcast News datasets prepared for the RT-03/04 shared tasks in MDE by the National Institute of Standards and Technology, U.S. Department of Commerce. We work with monologues recorded in language tests, and so we more closely follow the work of Lee & Glass (2012; L&G) because they trained and tested a system on monologue restaurant reviews. We choose to mimic L&G's work because firstly the results are interpretable (as opposed to machine learning with neural networks, e.g. Xu et al. (2014)). Moreover we can work with the relatively small learner datasets we have, and make use of them in an optimal fashion – *i.e.* using different corpora to train the separate components of the system where this brings performance improvement (sections 3.1 & 4.1). The modular architecture is appealing as it allows different model types to be combined – that is, models other than the finite state transducers introduced below, even though we do not do so in this work.

3.1 System architecture

The design of L&G's system involves a combination of local constraints containing prosodic and language model information, and global constraints of SU-length. One insight from related work is that a tagging approach to the problem only considers local information: if the search space is constrained between a minimum and maximum SU-length we can instead compute the likelihood of an SU boundary, denoted <$break$>, at each inter-token interval (Matusov et al., 2006). The models are implemented with finite-state transducers and combined in a log-linear fashion, such that the problem becomes not one of tagging but instead of finding the best path through the internal SU structures of our transcripts. Figure 1 gives an overview of our system architecture which is in spirit inspired by L&G though it differs in the detail, as discussed below. It is a three-part system using several probability sources (prosodic, lexical and SU-length) modelled as finite state transducers.

3.1.1 Prosodic model

As is the norm in the SU delimitation task, we build a prosodic model (PM) to predict SU boundaries. This move, and the feature types collected, reflect the assumption that speakers tend to indicate SU boundaries in consistent ways – by pausing before starting a new SU, by producing lengthened sounds in advance of an SU end, or by tell-tale discontinuities in pitch and volume levels either side of a <$break$>. Thus our feature choices are largely conventional, following the lead of Huang et al. (2006) as well as L&G (Table 1). This gives us a feature-set whose evolution can be traced back to the work of Shriberg and colleagues, with the obvious exception of turn-taking which is not available to us in monologues, though turns have been found to be a highly informative feature in dialogues (Shriberg et al., 2000).

For each token w_i we have two measures of pause duration (before and after), three phone duration features (the final phone, the sum of its vowels, and the longest phone), nine features for fundamental frequency (f0), and nine for energy – a total of 23 features for each token in the prosodic model. In the

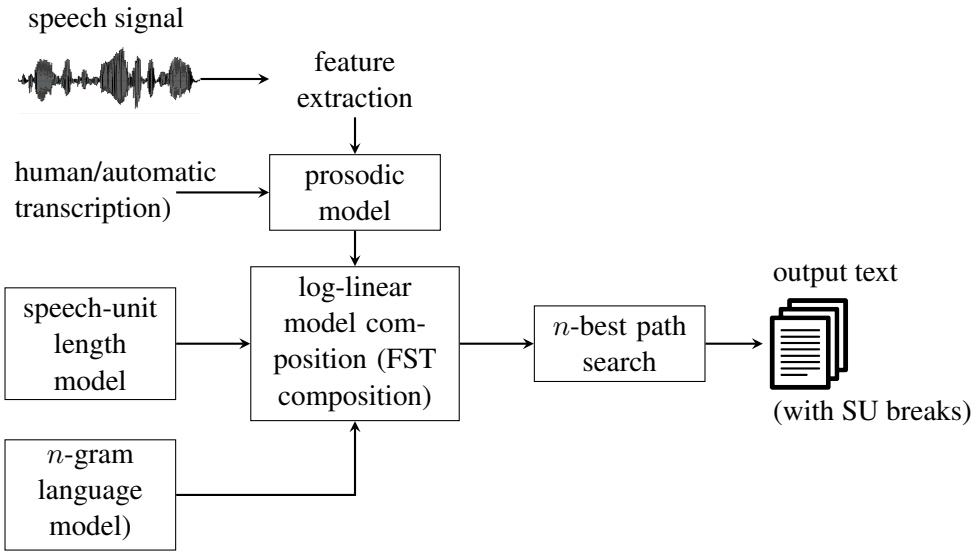

Figure 1: System diagram.

case of f0 and energy, the general task was to find maxima and minima in w_i and the following token w_{i+1}, or to calculate differences between the last frame in w_i and the first frame of w_{i+1}, or the minimum in the given recording, whilst also subtracting the speech-unit minimum from the mean across all frames in w_i, the start of w_{i+1}, and the mean of w_{i+1}.

All prosodic features were gathered automatically: transcriptions and sound files were force aligned using SPPAS (Bigi, 2012) before pause and phone durations were obtained with an R script (R Core Team, 2016), whilst fundamental frequency ('f0', measured in Hertz) and energy values (measured in decibels) were extracted in 10 millisecond frames using Praat's auto-correlated pitch and intensity tracking algorithms (Boersma and Weenink, 2016)[2]. As is conventional practice in signal processing – to normalise rapid, random changes in the signal – both f0 and energy values were smoothed using a five-point median filter and the 'robfilter' R package (Fried et al., 2014). Outlying phone durations were removed by filtering any tokens with a single phone posited to have been longer than two seconds, thereby excluding what were presumed to be gross alignment errors.

category	**features**	
pause duration	· pause before w_i · pause after w_i	
phone duration	· final phone in w_i · sum of vowel phones in w_i · longest phone in w_i	
f0	· max.f0 in w_i · min.f0 in w_i · max.f0 in w_{i+1} · min.f0 in w_{i+1}	· end of w_i − start of w_{i+1} · end of w_i − recording min.f0 · mean of w_i − recording min.f0 · start of w_{i+1} − recording min.f0 · mean of w_{i+1} − recording min.f0
energy	· per f0	

Table 1: List of prosodic features for the current word token (w_i).

This feature-set serves to capture the observed association between prosodic discontinuity and SU boundaries. That is, speakers tend to pause and lengthen SU-final tokens – hence the pause and phone duration features – and the pitch 'reset' between tokens either side of a boundary is likely to be more pronounced

[2]The Praat script was adapted from one written by Peggy Renwick, University of Georgia, for the BAAP workshop on 'methods for large-scale phonetic data analysis' held on 7 April 2014 in Oxford, U.K. We thank John Coleman, University of Oxford, for sharing it with us; `http://www.phon.ox.ac.uk/jcoleman/BAAP_workshop_info.html`.

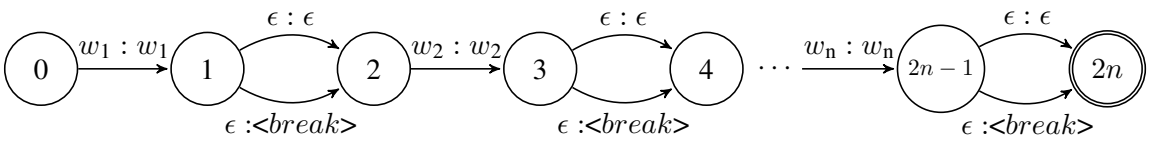

Figure 2: Prosodic model of an input string of length n.

than elsewhere: hence the focus on differences with the following token and the SU minimum (Shriberg et al., 2000). Energy features were not used by Shriberg et al. (2000) for reasons of data quality, but they were introduced by L&G on the basis of work by Huang and Zweig (2002) and so we use them here.

Per L&G, we trained a support vector machine (SVM) classifier with an RBF kernel trained on our twenty-three features, and used this to predict the probabilities of SU <$break$> tokens between word tokens. As a novel development we trained a logistic regression (LR) classifier using the six most significant features. Prosodic models can be modelled using an FST with a chain-like structure (Figure 2). Each odd-numbered node has a single arc to represent a word token, and each even-numbered node has a pair of arcs to represent what happens between those word tokens: one arc emits an empty string (modelled as an ϵ token), the other a <$break$> token, with the probabilities of these arcs being taken from the SVM or LR classifier based on the prosodic features of the given word token w_i.

3.1.2 Language model

To model the probability of local word ordering, we constructed an n-gram language model (LM). We used the OpenGRM library (Roark et al., 2012) to build models from native speaker and learner corpora. OpenGRM n-gram models are cyclic weighted FSTs, with a unigram state representing the empty string, and proper n-gram prefixes represented as their own states, so that an n-gram ($w_1..w_n$) is represented as a transition from its prefix state ($w_1..w_{n-1}$) via a word arc (w_n). Language models of this type take the same word for both input and output on each transition.

OpenGRM n-gram models use Witten-Bell smoothing by default (Witten and Bell, 1991) and back-offs are modelled using ϵ (empty string) arcs which allow the model to transition to a lower-order n-gram ($w_2..w_n$) should no state for ($w_1..w_n$) exist. To model SU delimiting <$break$> tokens we include them explicitly in the positions they occur in the training corpora. We also include an <unk> token to reserve some probability mass for out-of-vocabulary words. In all experimental settings we build 4-gram LMs.

3.1.3 Speech-unit length model

The third and final source of probabilities comes from the gold-standard length of speech-units gathered from our training corpora. Again following L&G we fit a gamma distribution to a histogram of gold-standard SU lengths. We then use this distribution to obtain the probability of a <$break$> token occurring at any given length of speech-unit, in an SU length model (SLM). These probabilities are used in the construction of a cyclic FST, where each node represents an SU of a given length (Figure 3). The transitions can accept any non-<$break$> symbol (represented here as <w>) and at each length a <$break$> is modelled by a backwards arc that restarts the 'counter' for the next SU.

Our model differs from that of L&G as we use a simple <$break$> or 'no break' probability at each length, whereas their model uses $P(length = n)$ or $P(length > n)$. As per their model we set a hard upper length L, which is corpus-dependent. We insert an SU delimiting <$break$> token when this length is reached.

3.1.4 Model composition

Across all models, probabilities are encoded as negative logs. When the models are composed, these weights, or penalties, are summed together. To find the most probable route through the combined FST we search for the combined path with the smallest weight. Since the non-<$break$> characters are accepted and emitted unchanged at each stage, it is the <$break$> tokens, or lack of them modelled as ϵ, that separate the various paths through the FST.

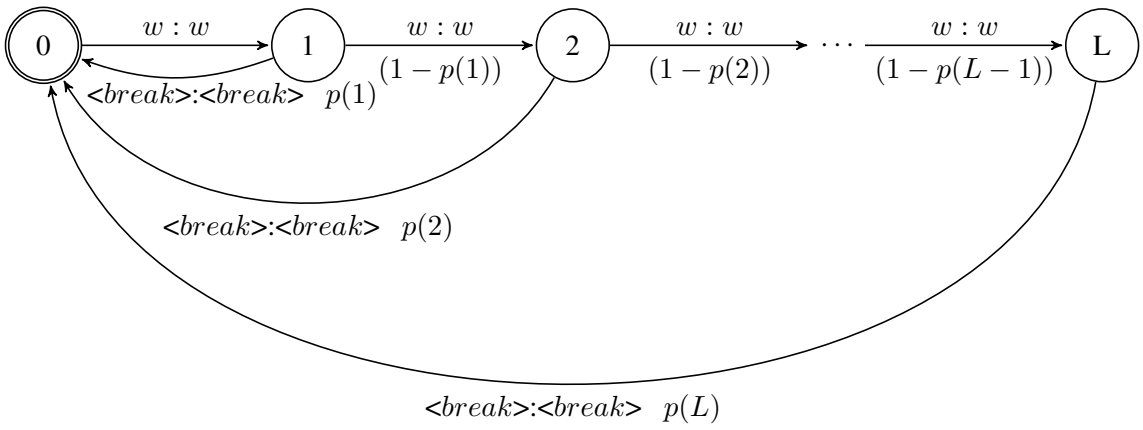

Figure 3: Speech-unit length model.

4 Experiments

The purpose of the system described above is to automatically assign SU-delimiting <$break$> tokens to transcripts of speech, based on a combination of prosodic features from the associated sound-file and probabilities from n-gram language and SU-length models.

4.1 Data

We work with a corpus of spoken learner English containing recordings of spontaneous speech from Business Language Testing Service (BULATS; `http://www.bulats.org`) oral exams. These feature 223 speakers whose first language (L1) is Gujarati – 96 male, 127 female, aged 14-50 (mean 25, median 24). Every recording has been graded by at least two expert assessors contracted by Cambridge English Language Assessment. This provides an average score for each learner which places them on the CEFR scale[3] from A1 ('beginner') to C2 ('mastery') via A2, B1, B2, C1 in increasing order of proficiency. Table 2 shows the distribution of learner CEFR levels in our BULATS corpus. It is apparent that the distribution of candidates and recordings is not equal, but there is no need for equivalence in this regard, as we seek only a representative sample of learners taking English exams. Note that there is only one candidate at the very highest C2 level, and therefore the corpus is very much a *learner* corpus mainly up to the C1 'advanced' level, rather than a corpus of Indian English learned as an L1.

CEFR level	Candidates	Recordings	Tokens
A1	33	209	6475
A2	44	288	10,597
B1	45	300	15,177
B2	44	304	17,921
C1	44	305	19,512
C2	1	6	383
Total	211	1412	70,065

Table 2: BULATS Spoken Learner Corpus.

Candidates were required to produce a monologue of twenty to sixty seconds on prompted business-related topics. Each recording was transcribed by two different crowdworkers via Amazon Mechanical Turk. Crowdworkers segmented the transcripts using punctuation, with full-stops (periods) indicating SU breaks. The two transcripts were then combined into a single version using the method described by van Dalen et al. (2015), which builds a network out of the two transcripts and uses an automatic speech

[3] 'Common European Framework of Reference for Languages' `http://www.cambridgeenglish.org/exams/cefr`

recogniser to identify an optimal path through it. Evaluating the quality of combined crowdsourced transcriptions, van Dalen et al. (2015) report a word error rate of 28.1% on another set of BULATS recordings. Inevitably, we pass word errors on through the pipeline described below, but as a method for the transcription of speech it at least gives us immediate access to large amounts of data. Phonetic transcriptions were then force-aligned with the recordings using the Hidden Markov Model Toolkit[4]. We found an error rate of 30% on phonetic alignments on a sample of the BULATS corpus, although such evaluation is not straightforward and a larger-scale exercise is needed in future work.

Transcribers indicated SU boundaries with full-stops and were asked to include all non-English words, partial words, filled pauses and repetition. The dataset features more than two hundred speakers, fourteen hundred recordings, and seventy thousand word tokens (Table 2). This is a small corpus by modern standards, and yet it is much larger than L&G's, which was 13.2k tokens. We trained a prosodic model as described in section 3.1.1 on a 90% set of 63k tokens and 2139 SUs identified by crowdworkers; the test set contains 7k tokens and 261 SUs.

We trained several 4-gram LMs (§3.1.2): firstly the learner English BULATS training set. This is only a small set of 63k tokens, and so we built a model of learner English based on a larger set of written exams from the Cambridge Learner Corpus (Nicholls, 2003), containing 766k tokens. Finally, we also trained a language model of native speaker transcripts, taken from the Switchboard Corpus of unscripted telephone conversations (Godfrey and Holliman, 1993), containing 217k tokens. Neither of the larger LMs are exactly apt for our learner monologues, in the sense that one is written language and the other is native speaker dialogue, but their greater size mitigates somewhat for the mismatch. We report in section 4.3 how all three LMs perform.

We experimented with different SLM models to little effect, and so for all experiments reported below we used an SLM model trained on our BULATS training set (§3.1.3). Maximum SU length was set to the longest SU found in the training corpus, which in this case was 185 tokens (*i.e.* if an SU reaches 185 tokens, the 186th token is a forced <$break$>). Figure 4 illustrates the distribution of SU lengths in our BULATS corpus.

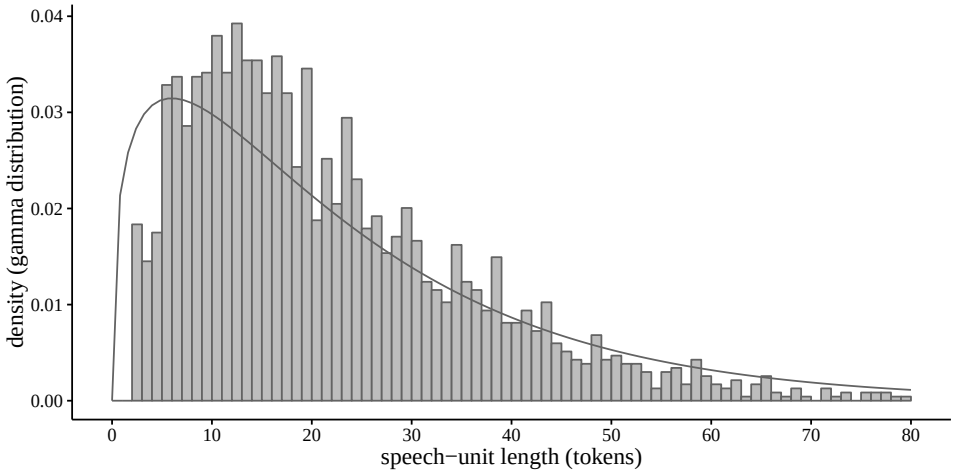

Figure 4: Density plot of speech-unit lengths in the BULATS learner corpus.

4.2 Evaluation

Lee and Glass (2012) report a BLEU-like score inspired by machine translation evaluation metrics (Papineni et al., 2002). L&G's modification is to only include n-grams consisting of hypothesised <$break$> tokens, which we have interpreted to mean any n-gram of size 1 to n with at least one <$break$> token within it. Their best performing system, with the PM weighted by 2.5, achieved a BLEU-like score of 0.56 on crowdsourced transcriptions. Unlike L&G we also report precision, recall and F-measure with

[4] `http://htk.eng.cam.ac.uk`

respect to the position of <$break$> tokens in the crowdsourced transcripts, because the BLEU-like score is a measure of *approximation* to the target inferred from the local context where <$break$> tokens occur, whereas F-measure with a single reference text is an exact score (although subject to the annotators' preferences).

These metrics very much rely on the idea of a 'gold-standard'. But as with so many language annotation tasks, the position of <$break$> tokens is highly subjective. Any downstream task requires that automatically identified SUs are useful in the sense that they can be passed to NLP tools for further inferential and processing tasks. Therefore we also report measures of SU quality, firstly with parse likelihoods from the RASP system (Briscoe et al., 2006), normalised by the number of nodes in the parse tree and averaged over each transcript. This gives us an idea of how useful the SUs are to downstream NLP tasks. Secondly we report perplexity scores for each transcript (with <$break$> tokens) obtained using the CMU-CAM toolkit (Clarkson and Rosenfeld, 1997) against a model of spoken learner English. This gives us an idea of how useful the SUs are for learner feedback in CALL systems.

Both of our new measures rely on language models and the idea of linguistic truth to some extent, but they are probabilistic, generalised over many SUs rather than the one-to-one comparison used in BLEU and precision/recall metrics. In both cases we can compare the scores for our hypothesised <$break$> tokens against the reference transcriptions and assess whether our SU delimitation system produces outputs we can work with in downstream CALL and NLP tasks. It is a matter for future work to investigate segmentation similarity metrics such as 'boundary edit distance' proposed by Fournier (2013)[5].

4.3 Results

We report modified-BLEU, precision and recall, parse likelihoods and perplexity scores (means and standard deviations) for our SU outputs in the BULATS test-set in various experimental configurations (Table 3). We tested several configurations of model type, weighting and combination, and evaluate our outputs from a held-out set of the BULATS corpus using a modified BLEU-score à la Lee and Glass (2012), information retrieval evaluation methods (IR, *i.e.* p, r, F), parse likelihoods, and perplexities.

The best performing set-up (f) features the BULATS-trained PM weighted by a factor of 5 and by recall, with a logistic regression classifier trained on the top six features. This is combined with the 4-gram Switchboard LM and BULATS SLM. Other configurations are shown for comparison. These include a baseline configuration with an unweighted SVM-based PM and learner LM/SLM (a), the same system with an LR-based PM trained on the top six features (b), and the uppermost weighting L&G apply to the PM of 2.5 (c). Both set-ups offer marked improvements on the baseline. Furthermore, a weighting of 5 on the PM offers an improved F-measure, though a reduced BLEU-like score (d). Finally, we show the performance of alternative LMs, constructed from the written learner Cambridge Learner Corpus (e) and native speaker Switchboard Corpus (f). The Switchboard LM offers a performance gain compared to the CLC, indicating that native speaker 4-grams model our test data more closely than learner writing, despite its smaller size.

We see in Table 3 that there are only small differences in BLEU-like scores, above the L&G-like baseline, in (b) to (f). This indicates that the systems are inserting <$break$> tokens in appropriate positions, given the training data, even if they are not precisely correct compared to the gold standard. Precision (p) and recall (r) confirm the differences in this regard, and show that adding weight to the PM greatly improves recall (*cf.* (b) and (c)..(f)). Increasing this weight from the 2.5 used by L&G to 5 again improves p and r (*cf.* (c) and (d)..(f)). Finally, using the larger LMs from other domains (CLC – the Cambridge Learner Corpus of written exams, and SWB – the Switchboard Corpus of telephone dialogues) leads to the most accurate SU delimitation compared to our gold standard annotations.

Our alternative measures – parse likelihood and perplexity – indicate that the hypothesised SUs score similarly across the configurations. That is, parse scores are slightly down on the gold-standard (these are negative logs, so closer to zero is more probable), whilst perplexity scores are more noticeably down

[5]We thank reviewer 1 for pointing us to this line of work. Having approached the topic from a speech engineering perspective, we were only aware of information retrieval metrics (precision, recall, *etc*) being used for evaluation, but there turns out to be a research literature in computational linguistics looking at more subtle evaluations to give credit for 'near misses' in light of the fact that annotators "frequently disagree upon the exact position of boundaries" (Artstein and Poesio, 2008).

BULATS PM	**LM**	**SLM**	**BLEU-like**	***p***	***r***	***F***	**parse likelihood mean (st.dev.)**	**perplexity mean (st.dev.)**
gold	.	.	.	.	.	.	-1.56 (0.52)	23.6 (39.9)
(a) PM_{SVM}	BULATS	BULATS	0.51	0.409	0.582	0.48	-1.58 (0.52)	35.9 (49.6)
(b) $PM_{r.6LR}$	BULATS	BULATS	0.75	0.64	0.5	0.56	-1.6 (0.53)	40.2 (60.7)
(c) 2.5*$PM_{r.6LR}$	BULATS	BULATS	0.75	0.639	0.617	0.628	-1.59 (0.54)	39.7 (60.8)
(d) 5*$PM_{r.6LR}$	BULATS	BULATS	0.74	0.653	0.686	0.669	-1.57 (0.54)	37.9 (58.7)
(e) 5*$PM_{r.6LR}$	CLC	BULATS	0.74	0.653	0.693	0.673	-1.59 (0.53)	39.2 (61.1)
(f) 5*$PM_{r.6LR}$	SWB	BULATS	0.74	0.656	0.693	0.674	-1.59 (0.54)	38.5 (60.4)

Table 3: Speech-unit delimiter output evaluation (PM: prosodic model; LM: language model; SLM: speech-unit model).

(the lower the score, the better the LM models the input) compared to gold, and the standard deviations indicate more variance within the output SU perplexities. This outcome reflects the error rate indicated by our F-measures – even though we outperform the state-of-the-art, we still have room for improvement before we can be sure that the SUs are entirely useful for downstream CALL tasks. However, the perplexity scores are low across the board and it does not seem therefore that the outputs are unfeasible. Since the parse likelihoods are similar to the gold standard, it seems that the outputs are syntactically feasible, which is promising for downstream NLP tasks – a matter we will fully evaluate in future work.

5 Discussion

Our best-performing configuration – 5*PM weighted by recall, with a logistic regression classifier trained on the top six features, and 4-gram Switchboard LM and SLM – compares favourably with the BLEU-like score of 0.56 reported by Lee and Glass (2012), and the state-of-the-art F-measure of 0.81 for SU delimitation in dialogues (Xu et al., 2014). L&G's optimal set up involves 2.5*PM based on an SVM classifier trained on their full set of twenty-three features (Table 1) and trigram LM/SLM.

It quickly became apparent to us that our prosodic model is strong (BLEU-like 0.728 alone), hence its greater weighting in our system. This was especially the case once we introduced a logistic regression (LR) classifier using the six most significant features to the PM, a refinement of L&G's method in this regard. For L&G the LM makes a huge improvement to the performance of the PM alone (from 0.13 to 0.53), while the SLM brings a more modest BLEU-like increase (to 0.56). In our case, the addition of the LM makes only a little difference (BLEU-like 0.735) with the SLM contributing another small increase (0.743). This outcome, so markedly unlike L&G's, requires several caveats. First, our LM is constructed from a relatively small corpus, with only 217k tokens in the Switchboard Corpus we use, compared to the 12m token web reviews corpus used by L&G. Secondly, it could be that native speakers of Gujarati transfer certain prosodic features that map well to speech-units when speaking English. To reinforce our results, we need to extend the system to learners with other L1s, a matter for future investigation. On the other hand, if the results stand up to further scrutiny, it suggests that the PM alone offers a good level of performance for SU delimitation, a possibility that would be beneficial resource-wise, as automatic extraction of prosodic features from the speech signal is a more straightforward exercise than subsequent combination with LM and SLM probabilities.

Moreover, we emphasise that the status of the 'gold-standard' in speech-unit annotation is uncertain. Both precision/recall and BLEU-like metrics rely on this idea, and hence we introduce other measures which relate to the 'likeliness' of the proposed SUs for downstream NLP and CALL tasks. Parse likelihoods and perplexity scores do not directly compare the hypothesised transcript to a gold-standard, but rather indicate the probability of such sequences, with <$break$> tokens positioned as they are. On these measures our outputs are syntactically feasible though more perplexing to a language model, reflecting the error rate implied by our F-measures in the range 0.4-0.7. Furthermore, as shown by the differences in BLEU-like and IR-type comparisons (Table 3) across SU delimiting systems, it is apparent that across the weighted-PM configurations, a similar performance is reported in terms of BLEU-like score, even where IR scoring varies. This suggests – because BLEU-like scores reward the prediction of n-grams commensurate with those found in the training data – that the proposed SUs in these cases may be 'good enough' even if they do not exactly correlate to the gold standard. Future work using segmentation similarity measures which explicitly reward near-misses will allow us to further investigate this matter (Fournier, 2013).

6 Conclusion

We have presented our work on speech-unit delimitation, firstly affirming that 'speech-unit' is the appropriate term for language chunks in spoken language, secondly releasing open-source software[6] to train and run an automatic SU delimiter constructed in a modular fashion per the work of Lee and Glass (2012). Thirdly, we considered alternative evaluation metrics for SU delimitation, ones which make use of the probabilities emitted by parsers and perplexity scores from statistical language models. We report the performance of our SU delimiter in various configurations on a spoken learner corpus, both with our new metrics and established BLEU-like and IR-type scores. Our best performing configuration makes use of a highly weighted LR-based PM and native speaker LM, demonstrating what we find most advantageous about this architecture: that its modular nature allows training on various sources, which is advantageous as the learner corpora we are interested in tend to be small resources.

ASR output transcripts are unpunctuated, and therefore an automated SU delimiter allows those transcripts to be subdivided and passed on to downstream applications in usable ways. In our case, we require SUs which are useful for automated learner assessment and feedback in CALL systems. In future we will continue to experiment with system configurations, data sources, and feature-sets for our SU delimiter. In addition, a further method to investigate how useful the outputs are would be extrinsic evaluation with users of a CALL system, to further consider what makes a meaningful speech-unit.

Acknowledgements

This paper reports on research supported by Cambridge English, University of Cambridge. We thank Dr Nick Saville and members of Cambridge English, Prof Ted Briscoe and colleagues in the ALTA Institute for their advice and support, and Gladys Tyen and Dimitrios Alikaniotis at DTAL. We are grateful to the three anonymous reviewers for their constructive criticisms which prompted us to make several improvements to the paper. We also thank Ann Lee and James Glass for their helpful insights into the system behind their 2012 INTERSPEECH paper, as well as Brian Roark for help with OpenGRM.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for Computational Linguistics. *Computational Linguistics*, 34.

Brigitte Bigi. 2012. SPPAS: a tool for the phonetic segmentation of speech. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association.

Paul Boersma and David Weenink. 2016. Praat: doing phonetics by computer [Computer program]. Version 6.0.14.

[6] `http://github.com/rjm49/multistage_segmenter`

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP System. In *Proceedings of the COLING/ACL 2006 Interactive Presentations Session*. Association for Computational Linguistics.

Ronald Carter and Michael McCarthy. in press. Spoken Grammar: where are we and where are we going? *Applied Linguistics*. doi: 10.1093/applin/amu080 (requires access to the journal *Applied Linguistics*).

Philip Clarkson and Ronald Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings ESCA Eurospeech*.

Alain Deschamps. 1980. The syntactical distribution of pauses in English spoken as a second language by French students. In Hans Dechert and Manfred Raupach, editors, *Temporal Variables in Speech: studies in honor of Freida Goldman-Eissler*. Mouton, The Hague.

Mark Dingemanse and Simeon Floyd. 2014. Conversation across cultures. In N. J. Enfield, Paul Kockelman, and Jack Sidnell, editors, *The Cambridge Handbook of Linguistic Anthropology*. Cambridge University Press, Cambridge.

Benoit Favre, Dilek Hakkani-Tür, Slav Petrov, and Dan Klein. 2008. Efficient sentence segmentation using syntactic features. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*. Institute of Electrical and Electronics Engineers.

Pauline Foster, Alan Tonkyn, and Gillian Wigglesworth. 2000. Measuring spoken language: a unit for all reasons. *Applied Linguistics*, 21:354–375.

Chris Fournier. 2013. Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Roland Fried, Karen Schettlinger, and Matthias Borowski, 2014. *robfilter: Robust Time Series Filters*. R package version 4.1.

John Godfrey and Edward Holliman. 1993. Switchboard-1 Release 2 LDC97S62. Web Download.

Jing Huang and Geoffrey Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*.

Zhongqiang Huang, Lei Chen, and Mary Harper. 2006. An open source prosodic feature extraction tool. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association.

Jáchym Kolář, Elizabeth Shriberg, and Yang Liu. 2006. Using prosody for automatic sentence segmentation of multi-party meetings. In *Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD'06)*, Berlin, Heidelberg. Springer-Verlag.

Ann Lee and James Glass. 2012. Sentence detection using multiple annotations. In *Proceedings of INTERSPEECH 2012*. International Speech Communication Association.

Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2005. Using conditional random fields for sentence boundary detection in speech. In *Proceedings of the 43rd Annual Meeting of the ACL*. Association for Computational Linguistics.

Evgeny Matusov, Arne Mauser, and Hermann Ney. 2006. Automatic sentence segmentation and punctuation prediction for spoken language translation. In *Proceedings of the 3rd International Workshop on Spoken Language Translation (IWSLT)*. International Speech Communication Association.

Diane Nicholls. 2003. The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of the Corpus Linguistics 2003 conference; UCREL technical paper number 16*. Lancaster University.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

R Core Team. 2016. R: a language and environment for statistical computing.

Manfred Raupach. 1980. Temporal variables in first and second language production. In Hans Dechert and Manfred Raupach, editors, *Temporal Variables in Speech: studies in honor of Freida Goldman-Eissler*. Mouton, The Hague.

Brian Roark, Yang Liu, Mary Harper, Robin Stewart, Matthew Lease, Matthew Snover, Izhak Shafran, Bonnie Dorr, John Hale, Anna Krasnyanskaya, and Lisa Yung. 2006. Reranking for sentence boundary detection in conversational speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers.

Brian Roark, Richard Sproat, Cyril Allauzen, Michael Riley, Jeffrey Sorensen, and Terry Tai. 2012. The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics.

Benjamin Roberts and Kim Kirsner. 2000. Temporal cycles in speech production. *Language and Cognitive Processes*, 15:129–157.

Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32:127–154.

Andreas Stolcke and Elizabeth Shriberg. 1996. Automatic linguistic segmentation of conversational speech. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*. International Speech Communication Association.

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Mari Ostendorf, Dilek Hakkani, Madelaine Plauche, Gokhan Tür, and Yu Lu. 1998. Automatic Detection of Sentence Boundaries and Disfluencies Based on Recognized Words. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*. International Speech Communication Association.

Stephanie Strassel, 2003. *Simple metadata annotation specification, Version 5.0, Linguistic Data Consortium,* `https://catalog.ldc.upenn.edu/docs/LDC2004T12/SimpleMDE_V5.0.pdf`.

Marcus Tomalin, Mark J. F. Gales, X. Andrew Liu, Khe Chai Sim, Rohit Sinha, Lan Wang, Philip C. Woodland, and Kai Yu. 2007. Improving Speech Transcription for Mandarin-English Translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*.

Rogier van Dalen, Kate Knill, Pirros Tsiakoulis, and Mark Gales. 2015. Improving multiple-crowd-sourced transcriptions using a speech recogniser. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers.

Christopher Walker, Stephanie Strassel, Elizabeth Shriberg, Yang Liu, Jeremy Ang, and Haejoong Lee. 2005. RT-04 MDE Training Data Text/Annotations LDC2005T24, Linguistic Data Consortium, `http://catalog.ldc.upenn.edu/LDC2005T24`.

Ian Witten and Timothy Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37:1085–1094.

Chenglin Xu, Lei Xie, Guangpu Huang, Xiong Xiao, Eng Siong Chng, and Haizhou Li. 2014. A deep neural network approach for sentence boundary detection in broadcast news. In *Proceedings of INTERSPEECH 2014*. International Speech Communication Association.