INTERSPEECH 2013

Using Conversational Word Bursts in Spoken Term Detection

Justin Chiu, Alexander Rudnicky

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA. USA

jchiu1@andrew.cmu.edu, air@cs.cmu.edu

Abstract

We describe a language independent *word burst* feature based on the structure of conversational speech that can be used to improve spoken term detection (STD) performance. Word burst refers to a phenomenon in conversational speech in which particular content words tend to occur in close proximity of each other as a byproduct of the topic under discussion. To take advantage of bursts, we describe a rescoring procedure that can be applied to lattice and confusion network outputs to improve STD performance. This approach is particularly effective when acoustic models are built with limited training data (and ASR performance is relatively poor). We find that word bursts appear in the four languages we examined and that STD performance can be improved for three of them; the remaining language is agglutinative.

Index Terms: Spoken Term Detection, Word Burst, Conversational Speech

1. Introduction

The Spoken Term Detection (STD) task [1] involves detecting the presence and location of targets, consisting of a word or a sequence of words in an audio corpus. Work in this area [2, 3] has shown that better performance is achieved when using a two-step procedure, ASR followed by indexing and search (as opposed to direct search of the audio). With this approach, the performance will be highly dependent on the accuracy of recognition; the approach is particularly suited to cases where adequate training data are available for ASR. On the other hand, given a language with limited data, creating a high-quality recognizer becomes more difficult. We believe that overall system performance can be improved through the use of knowledge about human conversation (since, effectively, corpora of interest involve interaction between two or more humans). In this paper we present some work that exploits one feature of conversation, which is that it is for the most part organized around topics. A byproduct of this will be the reuse of particular words; we can therefore expect tokens of the same word to typically recur several times in proximity which we refer to as *word bursts*. Using this knowledge, we can improve the performance on Spoken Term Detection.

The idea we propose is similar to that of cache language models [4], first proposed in the early 1990s, but makes explicit use of structural properties of conversational speech. Cache-based language models [4] used a “cache component” that biases the model towards words that might be expected to be present for certain topics. The work described in this paper focuses on rescoring instances of particular words that are identified in an output lattice (specifically, a CNC) and does not make use of topic information.

A desirable characteristic of this technique is that it appears to be language-independent (conversations, it would seem, are similar across languages). We present our work on 4 languages (Cantonese, Pashto, Tagalog and Turkish) using output from the Janus [5] ASR system. The result indicates that, to differing degrees, word burst processing can improve STD performance on all these languages.

We review previous research in section 2. We introduce the word burst approach in section 3. The dataset will be described in section 4. Result will be reported in the section 5 and analysis in section 6, followed by a discussion.

2. Previous Work

Spoken Term Detection is a relatively recent topic in speech research. Prior to STD, research efforts in this domain centered on classical Information Retrieval techniques applied to speech transcripts. As a result, the effectiveness of the retrieval mostly depended on the accuracy of the transcription. Jonathan et al. [2] discovered a significant drawback of this approach, which is its inability to deal with out-of-vocabulary words. David et al. [3] proposed a different approach which uses the information from the lattice to do the indexing and search. It estimates word posteriors from the lattices and use them to compute a detection threshold that minimizes the expected value of a user-specific cost function. The system also accommodates out-of-vocabulary search terms by using approximating string matching on phonetic transcript. Both approaches are distinct Spoken Term Detection from traditional Information Retrieval search, in that the approaches are starting to rely on the information that is available only from speech recognition.

These approaches perform well for languages that have sufficient ASR training materials, e.g. English, Chinese or Arabic but less so on languages with limited resources. Such languages currently include Cantonese, Pashto, Tagalog and Turkish. To compensate for relatively poor recognition accuracy, we can try to incorporate other sources of knowledge to enhance its performance.

Cache-based language models [4,8,9] were originally proposed to reduce perplexity and to improve speech recognition performance by maintaining a “cache memory” of recently encountered language units. The intuition was “a word used in the recent past is much more likely to be used soon than either its overall frequency in the language or a 3g-gram model would suggest.” We make the same assumption for conversational speech, but apply it in a different way.

Since we are working on languages that lack sufficient resources, we do not expect to have resources such as extensive training data or (e.g.) POS taggers available. But we conjecture that the structure of conversations is similar across languages and we exploit this feature.

3. Approach

We define a *word burst* as a naturally-occurring temporally local cluster of words. When in a conversation that touches on specific topics, the content word within the same topic will tend to occur near each other. It follows those words of interest for STD will (on the whole) be such content words.

To focus on words of likely interest, we take an existing vocabulary and (limited) text resource and use it to define a stop list as the most frequent words in the available corpus; we experimented with lists that include 1–5% of the vocabulary. This needs to vary according to language. In an agglutinative language such as Turkish, there are many morphological variants and this required a longer stop-word list.

Table 1. *Content word window size / burst percentage.*

	10 sec	15 sec	20 sec	25 sec	30 sec
Cantonese	43.6%	48.4%	51.3%	53.2%	55.0%
Pashto	35.7%	40.2%	43.3%	45.7%	47.9%
Tagalog	40.7%	45.0%	48.0%	50.0%	51.6%
Turkish	35.4%	39.2%	41.4%	43.1%	44.4%

Satanjeev et al. [6] proposed using a window size of 20 sec to detect the topic state in meetings; we used this as a starting point. Table 1 shows the percentage of content word within burst windows. We exclude words from the top 1% and all singletons in the available corpus. As can be observed, content words (as defined) tend to occur in bursts.

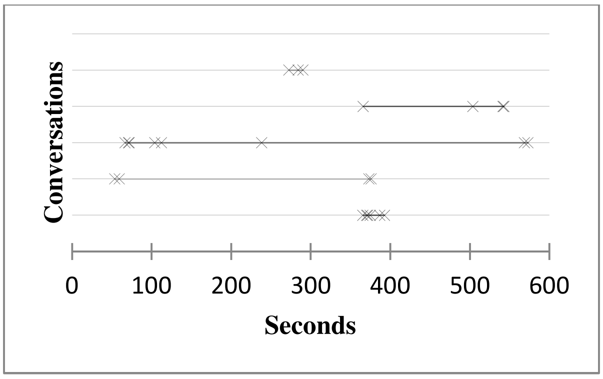

Figure 1: *Term incidence for Tagalog "magkano" (abscissa is time in sec; individual lines are separate conversations)*

Figure 1 provides a visual example, using the distribution of Tagalog term "*magkano*" in our data. For the graph, we can see the word have the tendency to occur in bursts. Our approach does confusion network rescoring to boost the scores of same-word hypotheses that occur in bursts, and penalizes words that do not. We use confusion networks to align hypotheses and to eliminate any spurious ones. The score-weighting formula is as follows:

$$R(t) = s(t) + b(t) \quad if\ \forall x \in window(t) \tag{1}$$

$$R(t) = p(L) * s(t) \quad if\ \forall x \notin window(t) \tag{2}$$

$$b(t) = \sum_i d(t, x_i) * \sum_i (d(t, x_i) * s(x_i)) \tag{3}$$

$$d(t, x_i) = 1 - (dis(t, x_i) / windowsize) \tag{4}$$

where $R(t)$ is the score for target word after rescoring, $s(t)$ is the raw score for term t, $b(t)$ is the boost term t will get from word burst, $p(L)$ is the language dependent penalty for each non-burst word, and $d(t, x_i)$ is the weight for word burst between target word t and it's word burst word x_i.

If there are burst words in the current word's burst window, we will boost current word's score by adding its burst term's score multiplied by a weight. The distance of the burst term and the current word decides the weight, which decays with distance. After all the summation, we multiply the boosting score by the sum of the weight for all the burst terms. The reason for this multiplication is to further boost the word that is frequently occurs in a small region. Even it was not placed in the high rank in the confusion network; it still has good chance of being the content word we need due to its repeated occurrence.

Three different parameters can be tuned in this approach:

- The size of the window to decide the word burst.
- The size of the stop word list
- The penalty for the word that does not have a burst.

Note that if the third parameter has been set to 1, the penalty function is deactivated. We will report the best parameters on the four languages we have.

4. Datasets

The data for the experiments are conversational (telephone) speech recorded in four different languages: Cantonese, Tagalog, Turkish and Pashto, languages that have become available through the IARPA BABEL program[1]. For each language, there are two different sizes of training data, 80 hours (FullLP) and 10 hours (LimitedLP).

Table 2. *Training data Lexicon size comparison.*

Language	**Setup**	**Lexicon Size**
Cantonese	FullLP	18,769
	LimitedLP	5,112
Pasto	FullLP	17,904
	LimitedLP	6,219
Tagalog	FullLP	21,098
	LimitedLP	5,565
Turkish	FullLP	38,849
	LimitedLP	10,173

Table 2 compares lexicon sizes in training data for the four languages. Turkish is known for being a morphologically rich language, which result in a much bigger lexicon size compare to other languages.

We had an additional 10 hours of development data for these four languages that can be used for tuning and testing. The experiments we describe below were carried out using a 5-fold cross validation (8 hours development data and 2 hours testing data) for the parameter tuning and for results.

[1] http://www.iarpa.gov/Programs/ia/Babel/babel.html

5. Results

Spoken Term Detection uses ATWV (Actual Term Weighted Value) for evaluation [1]. The formula for ATWV is as below:

$$TWV(\theta) = 1 - average\{P_{Miss}(term,\theta) + \beta \cdot P_{FA}(term,\theta)\} \quad (5)$$

$$\beta = \frac{C}{V}(\Pr_{term}^{-1} - 1) \quad (6)$$

where θ is the detection threshold.

Our cost/value ratio C/V is 0.1, thus the value lost by a false alarm is a tenth of the value lost for a miss. The prior probability of a term $\Pr_{term}$ is 10^{-4}.

Table 3. *ATWV Comparison on the full dev set.*

Language	Setup	Baseline	Rescore	Change
Cantonese	FullLP	0.322	0.320	-0.7%
	LimitedLP	0.103	0.110	+6%
Pashto	FullLP	0.214	0.215	+0.5%
	LimitedLP	0.095	0.114	+19%
Tagalog	FullLP	0.358	0.358	-0.3%
	LimitedLP	0.130	0.144	+11%
Turkish	FullLP	0.385	0.385	+0.1%
	LimitedLP	0.262	0.265	+1%

Table 3 shows the results for the FullLP and LimitedLP datasets. Word burst rescoring works better in the condition when it lacks enough training data. For FullLP data, it only produce insignificant difference to the Spoken Term Detection result, since the recognizer is robust enough and does not really need extra knowledge and information from the Word burst assumption. For the later analysis, we will focus on the Limited LP dataset.

Table 4. *Table of best parameters.*

Language	Window (sec)	Stop (%)	Penalty
Cantonese	12	4	0.1
Pashto	18	1	0.05
Tagalog	11	2	0.15
Turkish	9	8	1

We obtained our best parameters using the 5 fold cross-validation from the development data, which is shown in Table 4. The parameter with the best average result from 5 developing data in cross-validation is used for testing.

Table 5. *Result from cross validation test set.*

Language	Baseline	Rescore	Δ ATWV	FA
Cantonese	0.114	0.118	4%	-21%
Pashto	0.073	0.094	29%	-32%
Tagalog	0.143	0.159	11%	-21%
Turkish	0.241	0.244	2%	+4%

Table 5 shows the result from the cross-validation testing set. Aside from the change in the ATWV, we also show the change in false alarm rate after word burst rescoring. With the exception of Turkish we note substantial drops in false alarms. However, current uses of the ATWV metric in the community do not particularly emphasize false alarm errors. We also note that highly inflected languages do not respond well to rescoring as currently applied. We expect that this is taken into account, say by introducing suitable variants during burst identification, performance will improve.

6. Analysis

In this section, we will provide some observations on the effectiveness of the word burst rescoring approach on the Spoken Term Detection Task.

Improvement in the LimitedLP condition is always better than the improvement in the FullLP condition. This supports a common observation, which is that given sufficient data, basic modeling techniques can do very well. But of course sufficient data are not always available, and the introduction of additional information can compensate for its lack. A techniques such as word-burst rescoring is therefore of great use for languages that do not have extensive an accumulation of resources.

Table 6. *Query / Content word burst percentage.*

Language	Cantonese	Pashto	Tagalog	Turkish
Content	43.6%	35.7%	40.7%	35.4%
Query	41.2%	35.3%	36.8%	27.7%

Table 6 shows the comparison of burst word percentage of content word and the query words used in evaluation. The window size we pick here is 10 seconds, since it's close to the optimal window size that we identified in our parametric experiments (see Table 4). The query word was the query term provided for the development data. Table 6 shows that the distribution of query terms is similar to that of the content word that we identified in section 3, except for the agglutinative language. This difference also makes the optimal parameter for agglutinative language different from the other languages.

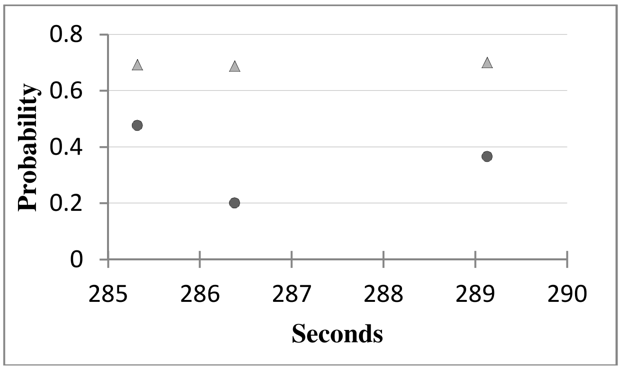

Figure 2: *Rescore effect on Tagalog "ganyan"*

Figure 2 shows how the word burst rescoring algorithm actually affects word probability. The abscissa indicates the time (in sec) of the word's start time. The ordinate indicates the probability of the Tagalog term "ganyan" in the confusion network. The lower (circle dot) line in the graph is the word's original probability in confusion network, and the upper (triangle dot) line is the rescored probability. The three "ganyan" words occurring over a small region boost each other's word scores. The STD system's threshold probability for word detection was set as 0.5, which means in the original confusion network all three occurrences of the word "ganyan" would have been labeled as not detected. After the rescoring, all of these occurrences' probability rise to around 0.7, which enables them to be found by STD system.

We computed the Word Error Rate (WER) on the best hypothesis (i. e. top-1) from the confusion network, but found that the difference between the baseline and the word burst rescored versions is less than 1%. Since we are focusing on boosting the score for the possible candidates to above the threshold for term detection, it may not matter that word burst does not result in improving the quality of the best hypothesis.

The tuned best parameters on four different languages reflect the linguistic differences and the conversational similarity among different languages. The best window length for word burst rescoring is around 10 secs. The Pashto best window size is somewhat larger than 10 secs, yet if we compare the ATWV difference between settings at the window sizes at 10 and 18, it is about 0.002, which is insignificant. This suggests that 10 sec is more likely to be a good default choice for identifying word bursts. Note that the query sets are necessarily different across different languages, which is another factor that might affect the parameter settings but which cannot be controlled.

The stop-word list size and penalty percentage should be discussed together. Pashto and Tagalog are similar in their alphabet-based format, so the stop word percentage and the penalty for non-burst word are not that different from each other. The Cantonese words require word segmentation, so the concept of stop word may be composed by different characters, which makes more words end up in the stop word list. Turkish is a morphologically rich language, which makes its parameter far different from the others. The lexicon size also reflects this phenomenon. The high percentage of stop word might be due to including stop words and their morphological variants. The optimized penalty for Turkish is 1, which means no penalty for the term that has no word burst. This is due to the morphologically variance in the language. It is not likely to have the same word re-occur in the conversation, since it might appear in another morphologically different form. As a result, penalty the word that does not have word burst in Turkish will just reduce accuracy, nevertheless boosting re-occurring identical words can still provide some benefit, as our results imply.

For Pashto and Tagalog, word burst rescoring provides significant improvement in ATWV, and Pashto is especially improved due to its weaker baseline. The improvement on Cantonese is slightly lower, which might be due to the quality of word segmentation. While Pashto and Tagalog have each word as their basic component in the confusion network, Cantonese uses the word level in the confusion network, which is more abstract than its basic component, characters. Still, with the non-word burst penalty function, the false alarm in term detection is significantly reduced in all three languages. Since the ATWV scoring metric has parameters that decide the cost ratio of false alarm and miss, if there is an application for which false alarms come at a greater cost for performance, the penalty applied to the words outside a word burst will be more valuable than currently. In the Turkish configuration, there is no penalty for the word outside the word burst. In this case, the false alarm actually increased, but the extra correct detections found by word burst rescoring still improves the ATWV score though the improvement is not as noticeable as compared to other languages.

7. Discussion

We described a language independent word burst phenomenon that exists in conversational speech and that can be used to improve performance in the STD task, particularly when language resources are otherwise limited. Nevertheless the rescoring process, as currently implemented, is still affected by language characteristics. In particular, word segmentation on the Cantonese data or word normalization in Turkish indicates that this is the case. Since word bursts are dependent on the nature of words, it implies that the processing needs to take into account language information to transform the input data into a stream of lexical units whose identify or similarity can be automatically assessed. We believe this will be the case.

Word burst rescoring might also be useful in dialogue systems, as an adjustment to state-specific language models [10].

8. Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0015. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

9. References

[1] Fiscus, Jonathan G., Jerome Ajot, John S. Garofolo, and George Doddington. "Results of the 2006 spoken term detection evaluation." In *Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational Speech,* pp. 51-55 (2007).

[2] Mamou, Jonathan, Bhuvana Ramabhadran, and Olivier Siohan. "Vocabulary independent spoken term detection." In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, vol. 23, no. 27, pp. 615-622 (2007).

[3] Miller, David RH, Michael Kleber, Chia-lin Kao, Owen Kimball, Thomas Colthurst, Stephen A. Lowe, Richard M. Schwartz, and Herbert Gish. "Rapid and accurate spoken term detection." In *Proc. Interspeech*, vol. 7, pp. 314-317 (2007).

[4] Kuhn, Roland, and Renato De Mori. "A cache-based natural language model for speech recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.12, no.6, pp. 570-583 (1990)

[5] Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ries, and Martin Westphal. "The karlsruhe-verbmobil speech recognition engine" In *Proc. of ICASSP-97* (1997).

[6] Banerjee, Satanjeev, and Alexander I. Rudnicky. "Using simple speech–based features to detect the state of a meeting and the roles of the meeting participants." *Proc. Of Interspeech, Jeju* (2004).

[7] Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr

Motl´ıcek, Yanmin Qian, Petr Schwarz, Jan Silovsky´, Georg Stemmer, Karel Vesely´ "The Kaldi speech recognition toolkit." In *Proc. Of ASRU* (2011).

[8] Jelinek, Fred, Bernard Merialdo, Salim Roukos, and Martin Strauss. "A dynamic language model for speech recognition." In *Proceedings of the DARPA Workshop on Speech and Natural Language*, pp. 293-295 (1991).

[9] Goodman, Joshua T. "A bit of progress in language modeling." *Computer Speech & Language,* vol.15, no.4, pp.403-434 (2001).

[10] Xu, Wei, and Alexander I. Rudnicky. "Language modeling for dialog system." *Proc. of ICSLP, Beijing* (2000).