Evaluating Prosody-Based Similarity Models for Information Retrieval

Steven D. Werner
University of Texas at El Paso
stevenwerner@acm.org

Nigel G. Ward
University of Texas at El Paso
nigelward@acm.org

ABSTRACT

Prosody is important in spoken language, and especially in dialog, but its utility for search in dialog archives has remained an open question. Using prosody-based measures of similarity, which also roughly correlate with dialog-activity similarity and topic similarity, we built support for "retrieve more like this" searches. Performance on the Similar Segments in Social Speech Task at MediaEval 2013 was well above baseline, showing the value of prosody for search.

1. INTRODUCTION

In most cases people searching in audio are probably not really interested in finding *words*. What people want is often information of some type, which may be characterized in part by dialog process or activity, for example recommending, answering a question, agreeing, forming a decision, telling life stories, making plans, hearing surprising statements, giving advice, explaining, and so on. In dialog, such activities and topics often are associated with characteristic prosodic features and patterns.

Our basic idea is to use a vector-space model of dialog activity, where each moment in time maps to a point in this space. This representation is obtained by applying Principal Component Analysis to 78 local prosodic features computed every 10ms calculated over a 6 second sliding window [2]. This feature set was choosen for simplicity of computation and for providing coverage of most of the prosodic aspects known to be most relevent for dialog. It resembles that used in [2], but with more volume features and fewer pitch features, more speaker features and fewer interlocutor features, and more narrow-window features close to the point of interest and fewer distant-context features. After PCA this gave 78 dimensions, ordered by how much of the variation they explained.

In previous work [4] we found that dialog timepoints which were proximal in this space tended to be similar not only in dialog activity but in topic as well. Here we extend this work to use better similarity models, and report positive results on a standard problem, namely the Similar Segments in Social Speech Task at MediaEval 2013 for which the task definition, data set, and evaulation metrics may be found in [5].

2. THE MODELS

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

The similar segments task is based on regions, but the dialog-space model is based on timepoints. For simplicity, the middle point of the query region is used as the characteristic point. The most similar (proximal) timepoints, across the entire corpus, are then found and returned, in order, as the ranked list of jump-in points.

We started with a similarity metric using simple Euclidean distance in the vector space, as described in [4]. However we observed that some of the dimensions seemed especially useful for the similarity computations and/or more revealing of dialog activities. We wanted our models to reflect this, with greater weights for such dimensions. Doing so sacrifices the distance metaphor, but is computationally similar. Specifically, for any two points in a dialog, x and y, we compute a weighted sum of their differences on the dimensions:

$$dissimilarity = \sum_{i=1}^{78} w_i |x_i - y_i| \quad (1)$$

First we tried this with uniform weights, giving the "dissim" results in the tables. We then tried optimized weights, trained using linear regression, where the target was a distance of 0 if x and y were similar, and 1 if they were not similar. Thus, for example, if two selected timepoints x and y both were located in regions that had been tagged as talk about "favorite movies," then x and y were counted as similar. If x and y shared no tags, they were counted as not similar. This is of course not ideal, since a point-pair might be similar even if not belonging to regions that were felt to be worth tagging. Sets of similar and non-similar timepoint-pairs were obtained by random sampling over the training set.

For sampling we experimented with various more restrictive definitions of similar. One type of constraint was to require agreement by at least some number of annotators in order to consider a timepoint pair as similar. For this the label names, were ignored (as always), and so the annotators might have considered the points to be similar in different ways entirely. The second type of constraint relied on the utility values ("weights") assigned by the annotators to their tags, higher the more informative and cohesive they thought the tagset was. For example, in one sampling we included only pairs whose connecting tag was rated 3, excluding those rated 0, 1, or 2 [3]. Requiring higher tagweights and more agreement gave higher-quality training data, but at the cost of reducing the quantity of similar point-pairs available to train with.

We also experimented with pruning the dimensions, using

model	naive prec.	raw recall	raw s.u.r	norm. s.u.r	norm. recall	F
Random	6%	23%	0.25	0.86	0.83	0.86
Expl.	16%	46%	0.43	1.49	1.67	1.50
Distance	3%	26%	0.21	0.74	0.96	0.76
Dissim.	4%	26%	0.22	0.76	0.96	0.78
all+	6%	31%	0.26	0.89	1.12	0.91
good+	6%	34%	0.27	0.94	1.25	0.97
all-p+	7%	32%	0.27	0.92	1.17	0.94
good-p+	7%	34%	0.28	0.96	1.24	0.98

Table 1: Performance on Training Set. all = trained using all training-data similarity sets; good = trained on only point-pairs which were in same-tagged regions according to at least three annotators; p = iterative-leave-one-out pruning applied to dimensions, + = only positively-weighted dimensions retained; s.u.r = speaker utility ratio.

model	naive prec.	raw recall	raw s.u.r	norm. s.u.r	norm. recall	F
Random	7%	11%	0.12	0.43	0.40	0.43
Expl.	9%	18%	0.29	1.00	0.67	0.95
Distance	3%	16%	0.12	0.41	0.57	0.42
Dissim.	6%	22%	0.17	0.58	0.81	0.60
all+	7%	28%	0.22	0.75	1.03	0.77
good+	6%	22%	0.17	0.60	0.81	0.61
all-p+	7%	30%	0.22	0.77	1.08	0.79
good-p+	7%	26%	0.20	0.69	0.93	0.71

Table 2: Performance on the Test Set, as above.

two feature selection methods. This was prompted by the observation that linear regression consistently gave negative weights to some of the dimensions, for example 67, which, when we listened to it, seemed to encode the difference between calm, indifferent speech and energetic explaining. The first method was to try to leave a dimension out of the model (set its weight to zero), and if that improved performance on a held-out subset of the training data, to drop it from the set. This was iterated, typically resulting in dropping about a third of the dimensions. The second approach was to simply drop any dimension to which regression assigned a negative weight.

3. RESULTS AND DISCUSSION

The tables show the results[1]. for the four models which performed best on the training set and four reference models: the baseline, where the jump in points for each query are randomly selected; a tagset-exploiting model, where jump in points are found by considering tags by other annotators with regions that overlap the query region; the Euclidean distance model; and a model based on uniform-weight dissimalarity, that is, like distance but using absolute-value instead of squared differences. We used the tagset-exploiting model as a likely upper bound on performance, as it is akin to how a second human might themselves perform the search task. For the best models, performance is far above baseline, showing that information retrieval can indeed benefit by using prosodic information.

These results are, however, weaker than those that can be obtained by using lexical features. Perhaps in this corpus topical similarity was more relevent then functional similarity, and perhaps lexical models are better for topic similarity. Thus prosodic models may still be of value, as is, for languages for which speech recognizers are not available or perform poorly. We further conjecture that the prosody is capturing dimensions of similarity not seen in lexical similarity, and therefore that a combined model could do even better. Exploring this is a priority for future research.

The effects of using higher quality training data varied with the testset: on the training set, using the good quality set gave the best performance, but on the test set the model trained using all the data performed best. Pruning was generally beneficial, with dropping dimensions with negative weights being the most useful, with some additional benefit from also selectively dropping dimensions.

Looking at robustness to changes in the data, the picture is clouded by the fact that the test set was harder, in terms of recall (because the target regions, like all regions in this set, tended to be shorter and thus harder to find). Nevertheless, on the training set the best model's performance was still far above baseline, showing a degree of generalizability.

Although the potential utility of prosody for search has been long discussed [1], and demonstrations of the relevance for prosody for inferring emotion and dialog acts are common, here we demonstrate, for the first time, that prosodic information, used by itself, is actually of value for search in audio archives.

[1]From the point of view of the competition, these results are all unofficial, since the authors, being also the competition organizers, had privileged access to the data.

4. ACKNOWLEDGMENTS

We thank the National Science Foundation for support via a REU supplement to Award IIS-0914868, and Olac Fuentes.

5. REFERENCES

[1] D. Hakkani-Tur, G. Tur, A. Stolcke, and E. E. Shriberg. Combining words and prosody for information extraction from speech. In *Proc. Eurospeech, vol. 5*, pages 1991–1994, 1999.

[2] N. G. Ward and A. Vega. A bottom-up exploration of the dimensions of dialog state in spoken interaction. In *13th Annual SIGdial Meeting on Discourse and Dialogue*, 2012.

[3] N. G. Ward and S. D. Werner. Data collection for the Similar Segments in Social Speech task. University of Texas at El Paso, Technical Report, UTEP-CS-13-58, 2013.

[4] N. G. Ward and S. D. Werner. Using dialog-activity similarity for spoken information retrieval. In *Interspeech*, 2013.

[5] N. G. Ward, S. D. Werner, D. G. Novick, T. Kawahara, E. E. Shriberg, L.-P. Morency, and C. Oertel. The similar segments in social speech task. In *MediaEval Workshop*, 2013.