Using Prosody to Find Mentions of Urgent Problems in Radio Broadcasts

Nigel G. Ward, James A. Jodoin, Anindita Nath, Olac Fuentes

University of Texas at El Paso

nigelward@acm.org, jajodoin@miners.utep.edu, anath@miners.utep.edu, ofuentes@utep.edu

Abstract

This paper examines whether prosodic information is usefully indicative of urgency and related attributes of situations in news broadcasts. We find that, in all 8 languages studied, prosody is informative. We also find some predictive value in cross-language modeling, suggesting the possibility of universal tendencies.

Index Terms: urgency, semantic functions, stance, low-resource languages, information retrieval, filtering, prioritization, cross-linguistic, universals

1. Scientific Motivations

Previous research has shown that prosody is informative regarding various linguistic and paralinguistic functions [1, 2], but most prosody research to date has focused on a limited set of functions. The first aim of this paper is to broaden the range of inquiry, as this may lead to further exploration of new kinds of prosodic function and a more complete picture of the ways in which prosody conveys information.

Previous research has also shown that for some functions the prosody has similarities across languages, including the prosody of focus, turn taking, topic structure and, most popularly, emotion, where it has been well demonstrated that some aspects can be detected at above-chance levels using models trained on data from other languages [3, 4, 5]. However the scope of these inquiries has again been quite limited. The second aim of this paper is to explore the possible universality of the prosodic indications of other kinds of functions

Previous research on prosody has mostly addressed functions where the contributions of prosody are strong, salient, and consistent. However even when prosody serves only as a weak, ambiguous signal, the ways in which it conveys information are still of interest. The third aim of this paper is to explore how best to model prosody when the connections to specific functions are variable and weak.

2. Practical Motivation

To effectively provide humanitarian assistance after a natural disaster, responders need to understand the situation and know where the needs are greatest [6]. Audio streams, such as radio news, can be valuable sources of information, but generally include also much that is irrelevant. For example, reporting during a forest fire may include not only reports on the extent of the blaze, current containment efforts, and possible further developments, but also discussions of forestry management policies that may have contributed to the fire, eyewitness accounts of how the fire has changed scenic vistas, and interviews with families whose vacation plans have been disrupted. While on-topic, such information is of little interest to those organizing firefighting and relief efforts. Moreover, even during a disaster, radio broadcasts still include entirely unrelated topics, such as sports, entertainment, and politics. Thus there is a need for ways to automatically identify only segments that contain information that mission planners are likely to find useful.

For many world languages, there exist systems which can help, including speech recognizers and tools for filtering and search. However, disasters may occur anywhere in the world, without warning, so we need means for effectively filtering news audio in any language, including low-resource languages.

This paper examines the potential utility of prosody for this purpose, motivated by prosody's unique value proposition, in two respects. First, since some aspects of prosody exhibit universal tendencies, we can hope that indications of urgency and the like may also have prosodic universals. Second, for any given language, the prosodic inventory is much smaller than the lexical inventory, so we can hope that with just a small amount of time with a native-speaker informant we can identify the prosodic properties used by that language to convey urgency and related meanings.

3. Related Research

While the use of prosody for detecting urgency and related attributes has not previously been studied, some related functions have been. Prosody has been observed to have a role, in English, in conveying such stances as whether a statement is presenting good news or bad news [7], whether it includes a positive assessment [8, 9, 10], and whether it is about something in the here-and-now or something about the past or a speculated future [11]. Prosody has been used for detecting action items in meetings [12]. From studies of the prosodic expression of urgency specifically [13, 14, 15], for warnings and alarms, for example in the context of a semi-autonomous car requesting the driver to resume manual control, we know that higher pitch and faster speaking rate correlate with increased perceptions of urgency. Urgency may also involve a break with the ongoing flow of discourse, and studies of participants in realtime gameplay and in multi-task situations suggest that high pitch, pitch rise, and breathy voice may signal this meaning [16, 11]. In earlier work we found that news stories that included a call for "immediate action," such as preparing for an incoming hurricane or an invitation to attend a tonight-only movie showing, could be detected to some extent from specific prosodic configurations spanning a few seconds [17]. We have also found prosody useful for identifying location mentions [18].

4. Task and Data

Our investigation used data and task descriptions provided by Darpa's Low Resource Languages for Emergent Incidents (Lorelei) program [6, 19, 20]. This program addresses the challenge of developing language processing solutions rapidly for any low resource language where a need arises, and has fostered numerous advances [21, 22, 23, 24, 25, 26, 27].

Lorelei also provided an evaluation scenario designed to simulate what might happen after an actual disaster. At some time an event (an earthquake, for example) occurs somewhere in the world. Soon thereafter, analysts identify the relevant language, pull from the archives a few hours of radio news data for that language, unannotated, and send it to the tech teams. After receipt of this data, at time t, a tech team has 24 hours to develop a system to infer attributes of stories in that language. Upon delivery, the system is immediately put into production. For evaluation purposes, it is judged on its ability to accurately process a new, unseen set of data from this language. In this way the tech teams are judged on their ability to rapidly develop a system to process a surprise language. There is later a second system delivery point, 7 days after t, and this is again evaluated on unseen data.

This is a low resource scenario in that the surprise language data was of only a few hours, and was provided without annotations. However, to make the scenario realistic, teams were allowed to consult with a native speaker of the language for a short period of time: 1 hour for the 24-hour checkpoint, and 5 hours for the 7-day checkpoint. This was of course too short to develop a large, annotated training set, so a major challenge was how to effectively use this limited time.

The specific tasks for the systems were to make discriminations in support of identifying the news stories that included "actionable" information. In the context of humanitarian assistance and disaster relief, for information to be actionable, it should 1) be relevant to some specific incident (in Lorelei, relating to one or more of 11 specific types, such as evacuation, infrastructure damage, or flooding), 2) be urgent, 3) be current, that is, not something purely in the past or future, 4) be unresolved, not yet satisfactorily addressed, and 5) include an explicit mention of a location [19]. Accordingly, in the 2018 Lorelei evaluations, systems were required to process news segments and characterize each by filling in a "situation frame" with accurate values for all five of these attributes, among others. Thus, for example, each story had to be classified as urgent or not urgent. Each story was a fragment of a news broadcast that had been divided, by hand, to be mostly one topic and no more than 2 minutes in length.

Before the evaluations, Lorelei provided data for six languages — Bengali, English, Indonesian, Thai, Tagalog and Zulu — including news broadcasts that were segmented, and native-speaker annotations of each segment for each of the five attributes. While these are not all low-resource languages, for our investigations we limited ourselves to what was provided, on the order of 10 hours and 804 to 1132 segments per language. This data enabled us to experiment with different models and estimate system performance before the actual testing on the surprise languages.

For this task both recall and precision are important: we want to find as many urgent stories as possible, without too many false alarms. We therefore evaluated performance in terms of the area under the ROC curve, that is the true-positive versus false-positive curve, for each of the five attributes. (The official Lorelei metrics were dependent on inference also of additional attributes, which in our case were the responsibility of partner teams, and so were not helpful for our purposes.) This area under the curve (AUC) is used in all the tables below, with the baseline being of course always 0.50.

5. Four Explorations

We tried various approaches to this problem: one that didn't work, and three that did.

5.1. Exploration 1: Weighted Counts of Exemplar-Matching Patches

Our first attempt focused on the need to develop models rapidly, using only a small amount of native-speaker time. In general, most work today that uses prosodic information for classifications like this relies blindly on big data. Here this would require an unrealistic amount of native-speaker time, so we decided to explor an alternative strategy: to use knowledge of the space of possible ways in which the languages of the world assign prosodic feature combinations to various functions: emotional, turn-taking, focus, stance-related, and so on. If we can identify both the universal tendencies and the possible range of these variations, we may be able to efficiently pin down all the prosody-meaning mappings for any given new language, given only tiny amounts of data and a small amount of native-informant time.

For the task at hand, this suggested a strategy of presenting to the native informants just a few dozen short audio clips, optimally selected to be maximally informative regarding the meaningful prosodic patterns of the surprise language. To select such points we considered two unsupervised approaches. Both involved covering the unlabeled data with overlapping 6-second patches, and computed 88 diverse time-spread prosodic features for each patch. We first considered k-Means clustering over all patches, and then soliciting judgments of audio clips near the cluster centers. However, based on previous studies [28, 17], we estimated that several hundred clusters would be needed, exceeding the tight time budget. We therefore instead used Principal Component Analysis to discover the important dimensions of prosodic variation [29, 11]. Based on previous experience, we know that the top dozen or so dimensions are often meaningful, and that examining some extreme values, both positive and negative, often suffices to characterize the meaning of a dimension.

We tried this method for English, Spanish, and Bengali. The results in each case looked promising, in that each dimension-side's most extreme (exemplar) timepoints were generally also extreme in some semantic sense, and in that several dimensions related to our task. As each dimension is just a set of loadings over the per-patch prosodic features, we were thus able to identify meaningful prosodic patterns, that is, meaningful temporal configurations of prosodic features.

The next step was to use this information to build a classifier. As this method ascribes values to patches, we needed a way to aggregate these to derive values for story-level attributes. We tried counts or sums for timepoints which exceeded various thresholds, for each dimension. We evaluated such features first by measuring correlations, which were often modestly good, and then by using them in simple linear regression models, which however gave poor performance. Despite trying various various thresholds and weighting schemes, and even trying oracle-based selection of dimensions, overall the performance was never consistently better than chance.

5.2. Exploration 2: Patchwise kNN

Our second approach was to apply a model developed earlier for inferring stances from prosody [17]. This classified each patch in a story by k Nearest Neighbors using distance-weighted information from the labels of training-data patches (all represented as described above), and then classified each story using the average of the estimates for its patches. This method had shown good performance for stances relevant to the current tasks, such as "bad implications," "unusual or surprising," and "prompting immediate action," so we expected it to work well here also.

We tested this model using 80:20 training/test splits, building a language-specific model for each of the six languages. The results were mostly above chance, with urgency detected best, with an average AUC of .66 across the six languages. While good, this was lower that the performance we had expected. One possible reason is that the data in our earlier studies was exclusively local news, and thus viscerally urgent both for the newsreader and for the audience, and therefore probably expressed with more consistent prosody. In contrast, the Lorelei data included more national broadcasts, with the situation being described often remote both to the newsreader and to the listening audience. Most disturbingly, the variability was very high, meaning that we could not rely on the method to perform well for a surprise language. Nevertheless we kept it in our toolbox and evaluated it again later, as discussed below.

5.3. Exploration 3: Multi-Language Models using Linear Regression over Story-Wide Features

The previous two approaches were based on modeling the details of local prosody over regions of a few seconds. There is, however, a poor match between such models and the Lorelei task metrics, which evaluate the ability to infer attributes at the level of news stories, not utterances or utterance sequences. We accordingly developed models which are much less sophisticated but better match the task, using features computed at the level of stories. Our first such exploration was to try multi-language modeling. Not having data for all human languages, this was not universal, but rather a family of 6 models, trained and tested in leave-one-out procedure.

We investigated 29 features. There were 3 metadata features: broadcast ID, segment ID (position of the story within the broadcast) and segment duration. For the latter two we took the log, as preliminary investigations showed this to be slightly more informative. The other features were prosodic: there were 13 base features — an intensity feature, 4 pitch features, 2 rate features, 2 articulatory-precision features, a pitch-energy-peak alignment feature [30], speaking fraction, voicing fraction, creakiness, and the voiced-unvoiced intensity ratio — for each of which we used the mean and the standard deviation. All features were track-normalized to reduce the effects of speaker variation [31].

We first examined correlations of these features with urgency. A little preliminary experimentation on the English data showed that the standard deviations were most informative when the features were computed over 50ms windows, except for voicing fraction, for which a 500ms window was best. Since we are interested in possible universals, we looked for features that correlated consistently, either positively or negatively, with urgency across all 6 languages. We found a few: urgency correlates negatively with lengthening and with the segment ID, and it correlates positively with story duration and with the standard deviations of pitch height, pitch narrowness, enunciation, and speaking fraction. While not what we had expected from previous utterance-level research, these correlations are easy to understand: intuitively stories involving urgent situations may come early in the broadcast and may last longer, people conveying urgent information may speak faster, and urgent news stories often involve multiple speakers with various mental states, and thus exhibit variety in certain aspects of speaking style.

	relevant	urgent	current	unres.	located
Bengali	.56	.69	.54	.14	.54
English	.59	.66	.62	.47	.60
Indonesian	.64	.60	.61	.55	.57
Tagalog	.61	.66	.65	.52	.62
Thai	.57	.61	.52	.59	.53
Zulu	.47	.62	.42	.52	.50
average	.57	.64	.56	.46	.56

Table 1: *AUC with Multi-Language Linear Regression Models (leave-one-language-out)*

	Original	Augmented
Support vector regression	0.581	**0.654**
Linear regression	**0.639**	0.641
Regression forest	0.559	0.611
K-nearest neighbors	0.567	0.570
Multi-layer perceptron	0.638	0.635

Table 2: *AUC with Multi-Language Linear Regression Models, Average Results on Urgency*

To determine which features were most powerful for urgency classification, we trained random forests and observed which features were most often at or near the root. Three were especially powerful: segment length, segment ID, and speaking rate. Wide pitch range and enunciation (precise articulation) were also often near the root.

More systematically, we ran a feature selection algorithm and found that 7 features could be discarded without significant penalty, including the two derived from the voiced-unvoiced intensity ratio. We used the remaining 17 features for the experiments below.

The results are seen in in Table 1. For all 6 languages, performance is generally above chance. Prosody seems less informative for the resolved/unresolved distinction, and less useful for Zulu. The former was likely due to the data imbalance: vanishingly few segments were tagged as resolved, both in the training and test data. Regarding the latter, the Zulu newsreaders in this dataset seemed unusually consistent in tone, with less of the expressive variation apparent in the other datasets. Nevertheless it is remarkable that for urgency in particular, in news broadcasts in six languages, we found consistently good performance, with a leave-one-language-out AUC of 0.64.

We then explored ways to do better, refining the approach in two ways: first, by trying more sophisticated classifiers, and second in addressing the unbalanced training data problem. Since each language has many more instances of the negative that the positive class (averaging only 14%), we added synthetic positive-class data. Each synthetic example was created by randomly choosing two examples from the minority class and obtaining their weighted average, with random weights in the [0,1] range. The results of testing on one language when training with the original or augmented datasets from all the others are shown in Table 2.

	relevant	urgent	current	unres.	located
Bengali	.57	.61	.57	.09	.56
English	.53	.51	.55	.56	.55
Indonesian	.70	.43	.62	.40	.64
Tagalog	.65	.60	.67	.43	.63
Thai	.69	.62	.67	.42	.72
Zulu	.59	.64	.59	.58	.51
average	.62	.57	.61	.41	.60

Table 3: *Language-Specific Linear Regression Model Results, AUC*

model, training data	Kinyarwanda	Sinhala
Other-Language Models		
linear regression	**.57**	**.74**
kNN	.51	.52
Language-Specific Models		
linear regression, 1 hr of labels	.51	.59
linear regression, 5 hrs of labels	**.57**	.71
kNN, 5 hours of labels	**.57**	.51

Table 4: *Urgency AUC, new languages*

We observe that the synthetic data consistently improves performance, although to varying degrees, and that support vector regression trained on the augmented data does best overall. We also observe that simple linear regression does quite well, even without augmenting the data. This was surprising, but may be due to the small size of the dataset or to the robustness and appropriateness of our features.

5.4. Exploration 4: Language-Specific Linear Regression Models over Story-Wide Features

We next trained language-specific models, using the same features. Five-fold cross validation gave the results in Table 3. Comparing with the results in Table 2 we observe that the language-specific models perform better for most attributes, but that the cross-language model is better for urgency.

Seeking to understand the limits of this model, we examined some of the stories for which the performance was worst. We noted sensitivity differences among speakers, as in one report from a quiet suburban town, where a morning's late garbage collection was described using prosody that other newsreaders would likely reserve for truly urgent problems. We also noted style differences, as in one report where a police dispatcher, apparently lacking public speaking experience, read a disaster-related announcement in a flat voice, without much prosodic marking at all.

6. New-Language Results

The 2018 Lorelei program evaluation provided data for two surprise languages, Sinhala and Kinyarwanda. Following the scenario, we measured how well our models and methods worked for these previously unseen languages. These were applied using either little or no language-specific data, according to the scenario. First, for the 24-hour checkpoint, we generated results from two multi-language models, each trained on the 6 development-set languages. We then spent one hour with a native speaker of Kinyarwanda and one with a native speaker of Sinhala to obtain urgency judgments on some of the segments, and used these to build our first language-specific models. We continued to obtain judgments, and after a total of 5 hours working with native informants for each language, obtained annotations for a total of 282 segments of Kinyarwanda and 270 of Sinhala. For the 7-day checkpoint, we generated results for language-specific models trained on all this data. For each of the models, we evaluated performance doing leave-one-broadcast-out experiments.

As seen in Table 4, for urgency the performance of all models exceeded baseline. Linear regression generally outperformed k Nearest Neighbors. The cross-language model again outperformed the language-specific models, which is not surprising given the low-resource situation. While likely to be useful for practical purposes, the predictions were far from reliable. Seeking to understand why, we examined some of the stories for which the linear regression models' predictions were worst. A general problem was that some stories were ranked by our models as highly urgent, although the native informant had not annotated them as urgent. Listening, we found that many of these included music. Our features were designed to work for speech input only, and this shows that they are not robust to music, and should only be applied after diarization has been done to detect and discard music.

7. Discussion and Research Directions

We find that prosodic information is indeed informative for detecting several attributes of news broadcasts, namely urgency, current relevance, relevance to some kind of disaster, and the inclusion of a location mention. In terms of practical value, while the performance is not outstanding, there is currently no other way to so quickly produce a system capable of helping prioritize segments for a low-resource language. We expect these techniques also to be useful for applications beyond disaster response, in other situations needing filtering, characterization or prioritizing of audio segments. While in many workflows prosodic information alone may not be adequate for practical purposes, there is good potential for use in combination with other sources of information, such as lexical, since prosody can be a partly independent source of information.

In terms of furthering our knowledge of how much of prosody is universal, we have found that there are prosodic tendencies in the expression of urgency and other attributes that are consistent across 8 languages. Future work should further examine the generality of this finding.

We find, further, that very simple modeling techniques, including linear regression, generally worked well, but this may reflect primarily the small size of the available data.

Probably the best direction for improving these results would be to model per-speaker variation. This would enable better normalization with respect to typical sensitivity and style, and thus likely much better performance.

8. Acknowledgements

We thank Alonso Granados for help evaluating the kNN model, Chunxi Liu for discussion, and DARPA for support under the LORELEI program, although no official endorsement can be inferred.

9. References

[1] B. Schuller, "Voice and speech analysis in search of states and traits," in *Computer Analysis of Human Behavior*, A. A. Salah and T. Gevers, Eds. Springer, 2011, pp. 227–253.

[2] J. Cole, "Prosody in context: a review," *Language, Cognition and Neuroscience*, vol. 30, pp. 1–31, 2015.

[3] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-Cultural Psychology*, vol. 32, pp. 76–91, 2001.

[4] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition," *Psychological Bulletin*, vol. 128, pp. 203–235, 2002.

[5] S. M. Feraru, D. Schuller *et al.*, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 125–131.

[6] DARPA, "Low resource languages for emergent incidents (LORELEI)," 2014, Solicitation Number DARPA-BAA-15-04.

[7] J. Freese and D. W. Maynard, "Prosodic features of bad news and good news in conversation," *Language in Society*, vol. 27, pp. 195–219, 1998.

[8] V. Freeman, G.-A. Levow, R. Wright, and M. Ostendorf, "Investigating the role of 'yeah' in stance-dense conversation," in *Interspeech*, 2015, pp. 3076–3080.

[9] N. G. Ward and J. A. Jodoin, "A prosodic configuration that conveys positive assessment in American English," in *International Congress of the Phonetic Sciences*, 2019.

[10] V. Freeman, "Prosodic features of stances in conversation," *Laboratory Phonology*, vol. 10, no. 1, pp. 1–20, 2019.

[11] N. G. Ward, *Prosodic Pattterns in English Conversation*. Cambridge University Press, 2019.

[12] F. Yang, G. Tur, and E. Shriberg, "Exploiting dialogue act tagging and prosodic information for action item identification," in *IEEE Acoustics, Speech and Signal Processing Conference*, 2008, pp. 4941–4944.

[13] P.-S. Jang, "Designing acoustic and non-acoustic parameters of synthesized speech warnings to control perceived urgency," *International Journal of Industrial Ergonomics*, vol. 37, pp. 213–223, 2007.

[14] P. Bazilinskyy and J. C. F. de Winter, "Analyzing crowdsourced ratings of speech-based take-over requests for automated driving," *Applied Ergonomics*, vol. 64, pp. 56–64, 2017.

[15] M. Unoki, M. Kawamura, M. Kobayashi, S. Kidani, and M. Akagi, "How the temporal amplitude envelope of speech contributes to urgency perception," in *Proceedings of the 23rd International Congress on Acoustics*, 2019, pp. 1739–1744.

[16] F. Yang, P. A. Heeman, and A. L. Kun, "An investigation of interruptions and resumptions in multi-tasking dialogues," *Computational Linguistics*, vol. 37, pp. 75–104, 2011.

[17] N. G. Ward, J. C. Carlson, and O. Fuentes, "Inferring stance in news broadcasts from prosodic feature configurations," *Computer Speech and Language*, vol. 50, pp. 85–104, 2018.

[18] G. Cervantes and N. G. Ward, "Using prosody to spot location mentions," in *Speech Prosody, submitted*, 2020.

[19] "NIST LoReHLT 2018 evaluation plan, version 1.0.4," 2018, National Institute of Standards and Technology.

[20] S. M. Strassel, A. Bies, and J. Tracey, "Situational awareness for low resource languages: the LORELEI situation frame annotation task." in *Workshop on Exploitation of Social Media for Emergency Relief and Preparedness, at the European Conference on Information Retrieval*, 2017, pp. 32–41.

[21] P. Papadopoulos, R. Travadi, C. Vaz, N. Malandrakis, U. Hermjakob, N. Pourdamghani, M. Pust, B. Zhang, X. Pan, D. Lu *et al.*, "Team ELISA system for DARPA LORELEI speech evaluation 2016," in *Interspeech*, 2017, pp. 2053–2057.

[22] N. Malandrakis, O. Glembek, and S. Narayanan, "Extracting situation frames from non-English speech: evaluation framework and pilot results," *Proceedings of Interspeech*, 2017.

[23] U. Hermjakob, Q. Li, D. Marcu, J. May, S. J. Mielke, N. Pourdamghani, M. Pust, X. Shi, K. Knight, T. Levinboim *et al.*, "Incident-driven machine translation and name tagging for low-resource languages," *Machine Translation*, vol. 32, pp. 59–89, 2018.

[24] P. Littell, T. Tian, R. Xu, Z. Sheikh, D. Mortensen, L. Levin, F. Tyers, H. Hayashi, G. Horwood, S. Sloto *et al.*, "The ARIEL-CMU situation frame detection pipeline for LoReHLT16: A model translation approach," *Machine Translation*, vol. 32, pp. 105–126, 2018.

[25] M. Wiesner, C. Liu, L. Ondel, C. Harman, V. Manohar, J. Trmal, Z. Huang, S. Khudanpur, and N. Dehak, "Automatic speech recognition and topic identification for almost-zero-resource languages," in *Interspeech*, 2018.

[26] M. Yuan, B. Van Durme, and J. L. Ying, "Multilingual anchoring: interactive topic modeling and alignment across languages," in *Advances in Neural Information Processing Systems*, 2018, pp. 8653–8663.

[27] C. Liu, M. Wiesner, S. Watanabe, C. Harman, J. Trmal, N. Dehak, and S. Khudanpur, "Low-resource contextual topic identification on speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 656–663.

[28] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Interspeech*, 2016, pp. 495–499.

[29] N. G. Ward, "Automatic discovery of simply-composable prosodic elements," in *Speech Prosody*, 2014, pp. 915–919.

[30] ——, "A corpus-based exploration of the functions of disaligned pitch peaks in American English dialog," in *Speech Prosody*, 2018, pp. 349–353.

[31] ——, "Midlevel prosodic features toolkit," 2017, https://github.com/nigelgward/midlevel.