Inferring Stance from Prosody

Nigel G. Ward [1,2], *Jason C. Carlson* [1], *Olac Fuentes* [1],
Diego Castan [3], *Elizbeth E. Shriberg* [3], *Andreas Tsiartas* [3]

[1]University of Texas at El Paso, USA
[2]Kyoto University, Japan
[3]SRI International

nigel@utep.edu, jccarlson@miners.utep.edu, ofuentes@utep.edu,
diego.castan@sri.com, elizabeth.shriberg@sri.com, andreas.tsiartas@sri.com

Abstract

Speech conveys many things beyond content, including aspects of stance and attitude that have not been much studied. Considering 14 aspects of stance as they occur in radio news stories, we investigated the extent to which they could be inferred from prosody. By using time-spread prosodic features and by aggregating local estimates, many aspects of stance were at least somewhat predictable, with results significantly better than chance for many stance aspects, including, across English, Mandarin and Turkish, good, typical, local, background, new information, and relevant to a large group.

Index Terms: information retrieval, filtering, attitudes, sentiment, broadcast news

1. Motivation

Most work on speech for information retrieval and filtering has focused on topic and content, with some attention paid to a few other facets — notably including emotion, sentiment, and dialog acts. However these do not exhaust the aspects that one might use for information retrieval [1, 2, 3, 4, 5]. In this paper we consider stance. In the social sciences, "stance" and related terms refer to a very broad set of feelings and behaviors [6, 7, 8, 9], including all the nuances and subtleties of attitudes and related functions that people display in the course of pursuing various communicative goals. While stance is potentially broader than sentiment [10], previous work on modeling stance has examined only a few aspects, such as polarity and strength of opinion [11, 12]; in this work we examine a larger set.

The practical motivation for this work is to support filtering and finding patterns in news broadcasts. In the Lorelei scenario [13], a mission planner needs to find information relevant to organizing a humanitarian intervention after a natural or anthropomorphic disaster. Given the large volume of news broadcasts and social media communications that may potentially be relevant, analysts and planners need tools to organize these to gain situational awareness and support planning. Relevant aspects of these items include attitudes towards situations and facts, evaluations of different actors, the novelty of the information, the magnitude of the disaster, whether an input is well-informed or speculative, and its immediacy in time and place to the disaster and relief needs. Crucially, planners often work with data in languages where ASR and MT tools are rudimentary. They also frequently need big-picture information, such as in which valley the chatter about flooding has a more here-and-now stance. They may seek information from statistics and tendencies across many items, even when the categorization of any specific item is not highly reliable.

Table 1 shows the 14 aspects of stance considered in this work. This list reflects several considerations: opinions by some Lorelei program participants regarding likely utility, non-redundancy to what might be accomplished by topic-based retrieval, commonality of occurrence in disaster-related news stories, and ability to be reliably annotated [14].

This paper reports experiments in automatically detecting these 14 aspects of stance, as they occur in radio news in three languages, from prosody.

2. Data

To investigate the manifestations of these stance aspects, we assembled three sets of radio news broadcasts. The American English set is 650 minutes taken from archive.org, consisting mostly of broadcasts from WMMB, KBND, and CHEV, but including others chosen to increase the coverage of disaster-related topics, including shootings, protests, earthquakes, floods, power outages, hurricanes, various storms, epidemics, and wildfires. The Mandarin data is the first 279 minutes of the KAZN subset of the Hub4 collection [15]. The Turkish data is the first 11 hours of the Bolt Turkish language pack, LDC2014E115.

Each news broadcast was divided into news segments, with topics like: weather, hockey, parenting, bicycle race, jazz festival, hospital donation, erosion, evacuation, highway closing, drug arrest, job fair, burglary, and so on. Segments varied in length from tens of seconds to a few minutes, except that for Turkish long stories were split into 2-minute segments. Each segment was annotated for the presence of each stance aspect as 0 (absent), 1 (weakly present) or 2 (strongly present). Each stance was labeled independently, thus a given segment could be labeled, for example, as both deplorable and praiseworthy, if it mentioned both a deplorable act and a praiseworthy one.

Annotation for each language was done by three native speakers working independently. Their agreement levels were measured with average pairwise weighted Kappa, giving partial credit (0.5) for close matches, for example, a rating of 2 by one annotator and a 1 by another. As seen in Table 1, interannotator agreement was excellent for some stances and poor for others, depending on the language. Assuming that stances are, ultimately, continuous-valued phenomena, we use the average of the three annotators' labels as the "true" value for that stance for that segment. Examining correlations, we find that these 14 stances are largely non-redundant; for example in English the most related pair, bad and good, correlated at only –0.59. Table 1 also shows that most stances are not uncommon, reflecting that, while newsreaders strive to be authoritative and objective, they also humanize the news [16]. The lists of broadcasts

1. Bad Implications - information with undesirable consequences, such as a raise in taxes, an approaching storm, or a flood (0.72, 33%), (0.55, 15%), (.36, 18%)
2. Good Implications - the opposite, such as a peace agreement, a good harvest, or nice weather (0.46, 34%), (0.45, 19%), (.35, 5%)
3. Deplorable Action - something bad done by someone or some organization (0.74, 14%), (0.37, 3%), (.60, 17%)
4. Praiseworthy Action - the opposite: something good done by someone or something (0.35, 14%), (0.46, 8%), (.36, 6%)
5. Controversial - something people do or could disagree about, such as a bold action by some person or group, or a new government policy (0.53, 4%), (0.56, 5%), (.40, 25%)
6. Factual Information - information presented as facts (0.25, 96%), (0.59, 94%), (.41, 46%)
7. Subjective Information - the opposite, such as opinions, either the presenter's or someone else's, or information reported skeptically or speculatively (0.36, 11%), (0.55, 46%), (.45, 39%)
8. Unusual or Surprising - something quirky, odd, or unexpected (0.22, 6%), (0.69, 7%), (.30, 18%)
9. Typical or Unsurprising - the opposite, something expected (0.81, 31%), (0.59, 9%), (.14, 11%)
10. Local - personally relevant to the listening audience, like local weather or close-by rioting (0.39, 80%), (0.66, 46%), (.05, 8%)
11. Prompting Immediate Action - something that may motivate the listening audience to do something, like take shelter from a storm or vote in today's election (0.69, 20%), (0.74, 32%), (.06, 5%)
12. Background - the opposite, information useful just as background, such as an explanation of the causes of a situation (0.47, 36%), (0.60, 23%), (.21, 29%)
13. New Information - new information or description of a recent development (0.30, 41%), (0.92, 48%), (.19, 8%)
14. Relevant to a Large Group - something affecting many people (0.47, 48%), (0.83, 21%), (.28, 17%)

Table 1: *Descriptions of each stance aspect, abbreviated from the annotation guidelines [14], and statistics for English, Mandarin, and Turkish respectively: (interannotator agreement, percent of news segments with that stance present (label = 1 or 2))*

and segments, and the annotations themselves, are available at http://www.cs.utep .edu/nigel/stance/ .

3. Model

For the automatic inference of stance, many sources of information might be used. In this study we choose to explore only prosodic features. This is for several reasons. First, the need for humanitarian interventions often arises in areas where the language spoken is "low-resource," in the sense that tools and resources such as speech recognition, dictionaries, and large corpora may not be available. Second, while vocabularies differ arbitrarily, there are across languages universal tendencies for some prosodic features to express certain things [17, 18]. Third, previous research suggests that many aspects of stance might be expressed more by prosody than by words [19, 20].

Our approach is based on the observation that regions (patches) that are similar prosodically are often similar also in the stances they express.

We focus on regions because news segments are heterogeneous. A stance, when present, is not necessarily expressed continuously throughout a news story; rather, it may be indicated mostly in a few specific regions. For example, in a news story containing the sentence *Two SQ constables are being credited with saving three people from a burning house in Rowdon*, the prosodic indications that this was "praiseworthy" are present more on the subject and predicate than on the village name, let alone on the subsequent descriptions of the fire's origin. Thus stance inference is different from the more familiar classification tasks where something — such as an emotion, state or trait [21] — is assumed to be broadly present across the input, either because it is a direct indication of a mental or physical state, or because each input is short.

Ideally we would use a model of the rhetorical and discourse structures of news to locate the most informative regions for any specific type of stance, but no current model is suitable [22, 23]. Accordingly we use an estimate-locally-then-aggregate method [24, 25]. Thus, for each stance and each segment, every patch in the segment contributes an independent estimate of the strength of that stance in that segment. Patches are offset every 100ms, both in the test data and in the training (reference) data. Depending on the length of the segment, there may thus be tens or thousands of these estimates. The overall estimate is the average of these patch-based estimates.

Of the many possible ways to estimate the stance for each patch, we chose a nearest-neighbors algorithm, for three reasons. First, this makes minimal assumptions about the distributions. Second, this can handle cases where the relevant information involves configurations of features, not just distributions of feature values [26, 27, 28]. Third, as an initial investigation, we wanted an interpretable method, so that we could examine and learn from successes and failures. We implement nearest neighbors straightforwardly. For each patch in the segment to classify, we find the k most similar patches in the reference data set. For each of these k neighbors, we then look up, in the annotations, how that stance was annotated in the segment it came from. For example the nearest neighbor of a patch in the middle of *snap their losing streak with a win against* was a patch in the middle of *partly sunny and a warm day*. Since the latter patch came from a segment labeled "local=2, good=2, new=2," this is evidence that the sports segment is also conveying something that is locally-relevant, good news, and new information. A reference patch is more relevant to the extent that it is more similar to the new patch, so each neighbor contributes with a weight inversely proportional to the squared distance. Weights are normalized, so that estimates are not affected by the local density or sparsity of neighbors. The neighbors are found by computing distances in an 88-dimensional vector space, where each dimension is given by the values of one prosodic feature. Figure 1 overviews the method.

For this work we used three feature sets developed for other tasks. All attempt to broadly characterize the prosody across a

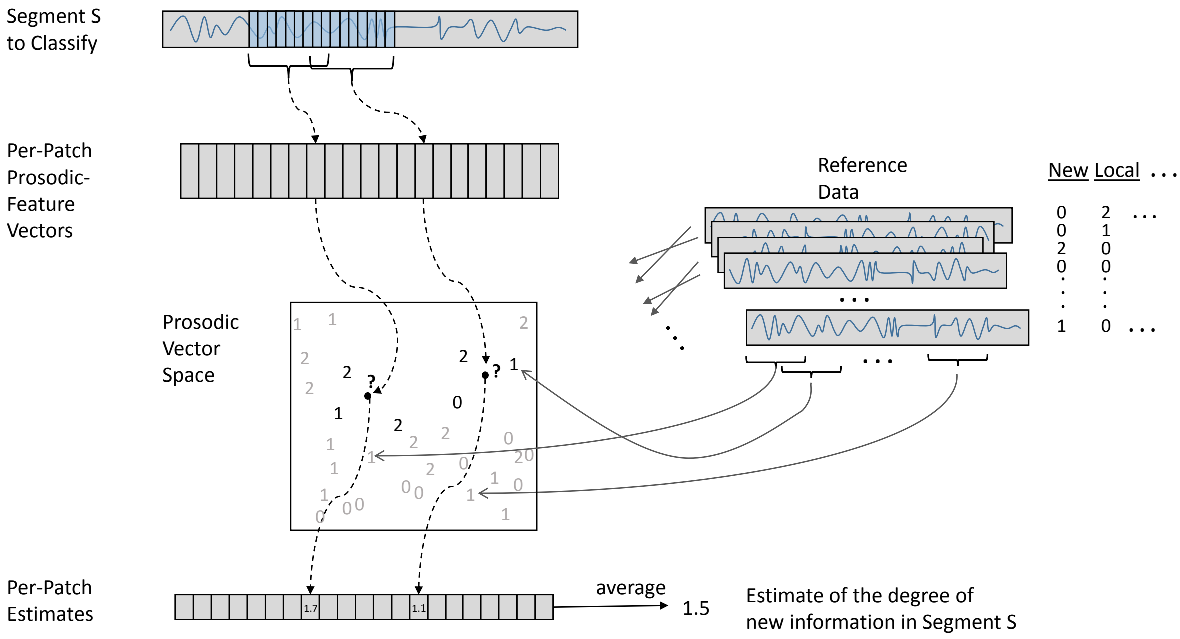

Figure 1: *Overview of the Method. Given a news segment S to classify, we take samples (patches) every 100 milliseconds, and represent each as an n-dimensional prosodic-feature vector. For each we find the 3 nearest neighbors, each representing a patch from the reference data, and each inheriting the stance annotations of the containing news segment. The stance of each patch is estimated using the values of its neighbors, and the overall stance of the news segment is estimated as the average of the estimates for each patch.*

patch using a large number of diverse features. The first was the Extended Geneva set of 88 features, a set designed for emotion recognition [29]. The other two were our own creations. Both of these include both wide-span features and features computed over narrower windows around a focal point within the patch. Neither includes features that are turn-, utterance-, word-, or syllable-aligned, so that all features can be everywhere-computable and robust. Most features are normalized to reduce dependence on the speaker and the recording conditions. In addition, each feature is z-normalized, across all audio tracks, so that, within each set, each feature contributes equally to the distance computations.

The second set was 88 prosodic features from the Midlevel suite [30], originally developed for language modeling and later extended for dialog-modeling and other purposes [31, 32]. These features tile a patch 6 seconds long, with a focal point in the middle of the patch. They include measures of intensity, pitch range, pitch height, speaking rate, and creakiness. The complete list is in [14], with the implementation details at [30], from which the code can also be downloaded.

The third set of features was 412 developed at USC and SRI for various purposes, including speech analytics and voice activity detection [33]. These features together tile a patch 4 seconds long, with a focal point at the end of the patch. The prosodic features are the averages and standard deviations of pitch and of intensity. In addition there are spectral features, namely long-term MFCCs, both raw and amplitude-normalized, and spectral-tilt features. There were also features for the percentage of voiced frames and for the energy ratio between voiced and unvoiced frames.

4. Experiments and Results

Our hypothesis is that prosody bears information useful for detecting what stance aspects are present in a news story.

A model should ideally assign to each segment the same label as the average of the human annotators, so our primary metric is the mean squared error. In this exploratory study, we are interested in determining whether prosody provides any information at all; thus for each stance aspect our baseline is the performance of a knowledge-free method: predict the average value across all segments in the reference set. This of course varied for each stance and language.

We used leave-one-out testing, that is, cross-validation at the segment level: for each segment and each stance, we predicted the value based on the annotations of that stance in other segments, across each entire dataset. In addition, for English, we did a known-speaker experiment, using the WMMB subset of the corpus. We also did cross-language experiments for English and Mandarin, where everything was the same, except that the nearest neighbors for reference were found in the data from the other language.

We tested our hypothesis by computing how close our prediction results were to the true values. We judged the predictor to be outperforming the baseline if its estimates were closer, $p < 0.05$ by a one-tailed matched-pairs t-test. Using $k = 3$ nearest neighbors, based on preliminary experiments, the results are as seen in Table 2. We had difficulty configuring the Geneva features to handle very short segments, so we lack directly comparable results, but on the other segments the Geneva set almost as well as the Midlevel set, for all three languages.

5. Discussion

Some stance aspects were predicted fairly well, indicating that prosody does indeed have value for predicting at least some stances. This was true for all three languages, for all three feature sets, and for both the large-data-multiple-speakers and modest-data-known-speaker conditions.

Overall, performance was better for Mandarin than for the

language	English	English	English	English	English	English	Mandarin	Turkish
test speaker(s)	mixed	mixed	mixed	mixed	mixed	known	mixed	mixed
reference minutes	-	-	650	650	650	70	279	672
reference segments	-	-	877	877	877	267	307	1038
feature set	-	-	midlevel	midlevel	SRI	midlevel	midlevel	midlevel
metric	baseline MSE	human MSE	predictor MSE	percent reduction	percent reduction	percent reduction	percent reduction	percent reduction
1 Bad	0.65	0.11	0.60	7%	18%	9%	**26%**	**18%**
2 Good	0.54	0.27	0.44	**17%**	**20%**	**16%**	**12%**	**12%**
3 Deplorable	0.39	0.07	0.36	8%	9%	3%	x 2%	**20%**
4 Praiseworthy	0.05	0.03	0.05	x 4%	x **-150%**	x -1%	**24%**	**16%**
5 Controversial	0.07	0.03	0.09	x -17%	x -17%	x -5%	10%	9%
6 Factual ...	0.07	0.06	0.08	x -5%	x 63%	x -2%	**36%**	**36%**
7 Subjective ...	0.14	0.08	0.17	-19%	0%	x -1%	**42%**	**18%**
8 Unusual ...	0.06	0.05	0.06	x 5%	x 29%	x -1%	11%	**19%**
9 Typical ...	0.77	0.09	0.52	**33%**	**47%**	**37%**	**73%**	**24%**
10 Local	0.37	0.27	0.23	**37%**	**42%**	x **13%**	**70%**	**19%**
11 Immediate ...	0.35	0.07	0.31	**13%**	**33%**	3%	**65%**	x 6%
12 Background	0.58	0.27	0.48	**16%**	**32%**	**18%**	**51%**	**6%**
13 New Information	0.39	0.31	0.30	**25%**	**28%**	**19%**	**75%**	**36%**
14 Large-Group ...	0.58	0.32	0.44	**25%**	**28%**	**21%**	**50%**	**27%**
average	0.36	0.15	0.29	18%	13%	9%	51%	22%

Table 2: *Performance in the various conditions. MSE: mean squared error. Percent Reduction: improvement in MSE over the baseline, as a percentage of the baseline. Bolding indicates statistically better than baseline by a two-tailed t-test; x indicates low variance (*< 0.10*), reflecting highly skewed priors.*

other two languages, both absolutely, as seen in the table, and in terms of closer approximating human performance. This may be because the KAZN data had more variety, including more acoustic variation between segments and more segments that were not simply read news but included spontaneous speech and dialog. In the cross-language experiments the performance was very poor: almost always below baseline. Clearly the prosodic reflections of stance differ greatly between English and Mandarin. Even in the best conditions, for most stance aspects the performance was well below human performance, indicating much potential for improvement.

One obvious factor related to poor performance was cases where the distribution for a stance was very unbalanced; in such cases, marked with x in Table 2, the dearth of reference-set diversity made the algorithm's task very hard.

Failure analysis revealed that some patches were less informative than others. For example, the prosody at one appositive-comma pause may strongly resemble the prosody at an appositive-comma pause in a different segment, regardless of the very different stances in the two segments overall. In general, there are times where prosody is being used to convey structure, not stance. This fact might be built into a model by using discriminative methods or by somehow using only patches expected to be informative.

All three feature sets performed above baseline, suggesting that this method is not over-sensitive to the exact features chosen. Performance was slightly weaker for the Midlevel 88-feature set, and experimenting with subsets revealed that its creaky-voice features and very fine-grained (50 ms) features were detracting from performance. Other experiments showed that further improvements could be obtained by including features for lengthening, delayed pitch peak, enunciation, and reduction, features now available in the Midlevel toolkit [30].

6. Significance and Future Work

This paper has explored the potential for using previously-unexplored aspects of stance for retrieval of spoken language. It presented a list of 14 aspects that are often relevant and frequently present in news stories, and showed that many of these are somewhat detectable automatically, from prosodic information alone. This serves as a first proof of concept of the idea of using stance for speech retrieval.

This paper also identified directions for improving performance. Future work should try more features, not only prosodic [34, 35, 36], but also, for some scenarios, spectral and lexical; try feature weighting and feature selection; and try discriminative, exemplar-based, and other models and machine-learning algorithms [37, 26]. For the latter, we should consider models where, rather than having each patch contribute equally, more weight is given to patches more likely to be informative, as estimated perhaps by some analog of inverse document frequency.

Also needing further study are the effects of data size and of speaker differences, and the issue of generality across speech genres, such as read news, interviews, speeches, dialog, and video soundtracks. Although the lack of commonality between the prosody-stance mappings of English and Mandarin suggests that expressions of stance are not universal, future work should examine generality within language families.

7. Acknowledgements

This work was supported in part by DARPA under the Lorelei program, by a Fulbright Award, and by the NSF through REU supplements to IIS-1449093. This work does not necessarily reflect the position of the Government, and no official endorsement should be inferred.

8. References

[1] M. Larson and G. J. F. Jones, "Spoken content retrieval: A survey of techniques and technologies," *Foundations and Trends in Information Retrieval*, vol. 5, no. 4-5, pp. 235–422, 2012.

[2] L.-S. Lee, J. Glass, H.-Y. Lee, and C.-A. Chan, "Spoken content retrieval: Beyond cascading speech recognition with text retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1389–1420, 2015.

[3] M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, S. Noorbaloochi, and S. Peters, "Detecting and summarizing action items in multi-party dialogue," in *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, 2007, pp. 200–211.

[4] M. Freedman, A. Baron, V. Punyakanok, and R. Weischedel, "Language use: what can it tell us?" in *49th Association for Computational Linguistics, Volume 2*, 2011, pp. 341–345.

[5] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *Intelligent Systems, IEEE*, vol. 28, pp. 46–53, 2013.

[6] J. Read and J. Carroll, "Annotating expressions of appraisal in English," *Language Resources and Evaluation*, vol. 46, pp. 421–447, 2012.

[7] O. Rambow and J. Wiebe, "Sentiment and belief: How to think about, represent, and annotate private states," in *Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL*, 2015.

[8] M. Chindamo, J. Allwood, and E. Ahlsen, "Some suggestions for the study of stance in communication," in *IEEE International Conference on Social Computing (SocialCom)*, 2012, pp. 617–622.

[9] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "SemEval-2016 task 6: Detecting stance in tweets," *Proceedings of SemEval*, vol. 16, 2016.

[10] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*. Springer, 2012, pp. 415–463.

[11] 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.

[12] 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.

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

[14] N. G. Ward, "Preliminaries to a study of stance in news broadcasts," University of Texas at El Paso, Department of Computer Science, Tech. Rep. UTEP-CS-16-66, 2016.

[15] S. Huang *et al.*, *Mandarin Broadcast News Speech (HUB4-NE)*. Linguistic Data Consortium, 1998, catalog No. LDC98S73, ISBN: 1-58563-125-6.

[16] C. Cotter, "Prosodic aspects of broadcast news register," in *19th Annual Meeting of the Berkeley Linguistics Society*, 1993, pp. 90–100.

[17] C. Gussenhoven, "Intonation and interpretation: phonetics and phonology," in *Speech Prosody*, 2002, pp. 47–57.

[18] J. Vaissiere, "Perception of intonation," in *The handbook of speech perception*, D. Pisoni and R. Remez, Eds. John Wiley & Sons, 2008, pp. 236–263.

[19] F. Mairesse, J. Poifroni, and G. Di Fabbrizio, "Can prosody inform sentiment analysis? Experiments on short spoken reviews," in *IEEE ICASSP*, 2012.

[20] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *ACL*, 2013, pp. 973–982.

[21] 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.

[22] P. C. F. Cardoso, M. Taboada, and T. A. S. Pardo, "On the contribution of discourse structure to topic segmentation," *Proceedings of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 92–96, 2013.

[23] S.-H. Liu, K.-Y. Chen, B. Chen, H.-M. Wang, H.-C. Yen, and W.-L. Hsu, "Positional language modeling for extractive broadcast news speech summarization," in *Interspeech*, 2015, pp. 2729–2733.

[24] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "A latent semantic model with convolutional-pooling structure for information retrieval," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, pp. 101–110.

[25] 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, San Francsico, USA*, 2016, pp. 495–499.

[26] Y. Kim and E. M. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3677–3681.

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

[28] L. Ferrer, E. Shriberg, S. Kajarekar, and K. Sonmez, "Parameterization of prosodic feature distributions for SVM modeling in speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, 2007, pp. IV:233–236.

[29] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, pp. 190–202, 2016.

[30] N. G. Ward, "Midlevel prosodic features toolkit," 2015, https://github.com/nigelgward/midlevel.

[31] N. G. Ward, A. Vega, and T. Baumann, "Prosodic and temporal features for language modeling for dialog," *Speech Communication*, vol. 54, pp. 161–174, 2011.

[32] 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.

[33] M. Van Segbroeck, A. Tsiartas, and S. Narayanan, "A robust frontend for VAD: Exploiting contextual, discriminative and spectral cues of human voice," in *Interspeech*, 2013.

[34] L. Ferrer, N. Scheffer, and E. Shriberg, "A comparison of approaches for modeling prosodic features in speaker recognition," in *IEEE Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4414–4417.

[35] M. Slaney, E. Shriberg, and J.-T. Huang, "Pitch-gesture modeling using subband autocorrelation change detection," in *Interspeech*, 2013, pp. 1911–1915.

[36] H. Arsikere, A. Sen, A. P. Prathosh, and V. Tyagi, "Novel acoustic features for automatic dialog-act tagging," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6105–6109.

[37] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *Interspeech*, 2009.