INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Exploring Dynamic Measures of Stance in Spoken Interaction

Gina-Anne Levow, Richard A. Wright

University of Washington,
Seattle, WA USA

levow,rawright@uw.edu

Abstract

Stance-taking, the expression of opinions or attitudes, informs the process of negotiation, argumentation, and decision-making. While receiving significant attention in text materials in work on the related areas of subjectivity and sentiment analysis, the expression of stance in speech remains less explored. Prior analysis of the acoustics of stance-expression in conversational speech has identified some significant differences across dimensions of stance-related behavior. However, that analysis, as in much prior work, relied on simple functionals of pitch, energy, and duration, including maxima, minima, means, and ranges. In contrast, the current work focuses on exploiting measures that capture the dynamics of the pitch and energy contour. We employ features based on subband autocorrelation measures of pitch change and variants of the modulation spectrum. Using a corpus of conversational speech manually annotated for dimensions of stance-taking, we demonstrate that these measures of pitch and energy dynamics can help to characterize and distinguish among stance-related behaviors in speech.

Index Terms: stance-taking, prosody, speech dynamics

1. Introduction

Stances, or a speaker's subjective attitudes or opinions about the topic of discussion [1, 2], are an integral part of activities involving collaboration, negotiation, and decision making. In automatic recognition research, stance is similar to sentiment and subjectivity, expressions of an internal mental or emotional "private state" [3]. Recognition research in these areas has grown rapidly following foundational work like [4, 5]. Generally, such work has relied on textual materials and annotated corpora, such as those described in [4, 5, 6]. Predominantly drawing on lexical and syntactic evidence, text-based approaches capitalize on well-formed sentences and complete thoughts; however, our focus is on stance-taking in spoken interactions, which involve ambiguous, fragmentary, or disfluent utterances. A much smaller amount of work has investigated issues of subjectivity, sentiment, or stance in speech, primarily by exploiting existing conversational dyadic ([7] in [8]) or multi-party meeting corpora ([9, 10, 11] in [12, 13, 14], respectively), small portions of which are annotated for elements of subjectivity such as agreement or arguing. Even with speech data, many approaches to automatic subjectivity recognition have leveraged mainly word or n-gram content [15], and efforts to incorporate prosodic information have yielded no significant improvement [16]. This is surprising, as stance-taking in speech harnesses channels of information not available in the textual content, including intonation, speaking rate, emphasis, and precision of articulation [17, 18, 19]. However, [13] found that annotators were better able to identify opinions, especially negative opinions, when they had access to audio recordings than when using transcripts alone.

Some recent work [20] has demonstrated the utility of simple prosodic and speaking style measures for stance strength and polarity recognition, both alone and in conjunction with textual word unigram features. Prosody and speaking style measures were found to improve recognition of stance with word information, except when manual punctuation information was provided. However, the prosodic measures employed in much of this work have been restricted to simple phrase-level functionals (e.g., maximum or mean) over standard pitch and intensity. At the same time, many of the channels for conveying stance information that have been identified rely on the dynamics of the signal itself. In this work, we focus on the use of prosodic measures that aim to directly capture the dynamics of the pitch and energy contours used to express stance in conversational speech, using a corpus of speech specifically targeting stance and manually annotated for stance-taking. We focus on measures that capture variants of the modulation spectrum of energy and subband correlation pitch change detection. We analyze the relationship between these dynamic measures and stance expression and employ these measures in automatic classification of stance strength, both alone and in combination with standard features. We demonstrate that these measures capture additional information that can serve to better discriminate among stance-related behaviors in speech.

2. Data: Corpus, Processing, & Annotation

The data for this study is drawn from the ATAROS corpus[1]. [18]. Unlike prior corpora employed for the study of subjectivity and stance in conversational speech [14, 16, 17], this corpus was designed explicitly to elicit high rates of stance-taking at different strengths, while controlling for regional dialect and recording conditions in a conversational setting. Pairs of participants performed a series of collaborative tasks involving negotiation and decision-making. In addition to two tasks designed to elicit stance-neutral first mentions of target lexical items, dyads completed three tasks designed to elicit increasing degrees of stance intensity and engagement through higher-stakes tasks. The **Inventory** task, designed to be low-stakes and to elicit lower levels of stance, required participants to discuss and agree on the arrangement of items in an imaginary superstore. The **Survival** task, designed to evoke moderate levels of stance and engagement, involved participants jointly deciding on items to salvage from a sinking ship to survive in a hostile environment. Finally, the **Budget** task, the highest stakes task, asked the participants to negotiate and decide on a series of cuts necessary to balance a hypothetical county budget. The effectiveness of this task design for manipulating stance and engagement was shown in [18]. Tasks averaged roughly 10 minutes in duration. Our analysis focuses on the two tasks at the extremes of the con-

[1] Available for research use from ttp://depts.washington.edu/phonlab/projects.html

http://dx.doi.org/10.21437/Interspeech.2017-1706

tinuum, with data drawn from the Inventory and Budget tasks.

The elicitation process yielded analyzable data from a total of thirty-one dyads. All participants were native speakers of English from the Pacific Northwest region of the United States (Washington, Oregon, Idaho). Dyads were matched and crossed for gender, and roughly matched for age within three broad age groups (18-32; 38-49; 60-75). For the analysis in this paper, we use a subset of the corpus that has been fully aligned, transcribed and annotated for stance-related behavior as described below. The subcorpus thus includes 21 dyads, with all but one dyad performing both the Inventory and Budget tasks. The 42 participants included 26 female and 16 male speakers.

All interactions were recorded in a sound-treated booth on individual channels using close-talking head-mounted microphones, sampled at 44.1kHz. All speech was manually transcribed following a simplified variant of the ICSI Meeting Recorder transcription guidelines [9], which uses conventional spelling, capitalization, and punctuation. The transcription was performed in Praat and coarsely aligned with the audio at the level of the 'spurt', a span of speech by a single speaker surrounded by at least 500ms of silence. Based on these coarse transcriptions, we performed automatic word- and phone-level forced alignment using the Penn Phonetics Laboratory Forced Aligner (p2fa; [21]).

The recordings were then manually annotated for a variety of stance-related behavior. Each spurt was annotated at the coarse level for stance strength and polarity. Stance strength was annotated for four levels: 0: no stance (back-channels, neutral facts); 1: weak stance (weak agreement, solicitation of opinion, weak opinions); 2: moderate stance (moderate agreement and attitude expression); 3: strong stance (particularly strong versions of the above). Stance polarity was labeled as: '+': positive (e.g., agreeing); '-": negative (e.g., disagreeing), and neutral. Annotation was performed based on listening to the spurt to be annotated in context, relative to the speaker's own style and taking into account both word context and prosody. After familiarizing themselves with the speaker's style, one annotator performed an initial labeling of the speaker within the task, using Praat [22] from playback and annotation. A second annotator reviewed and revised that annotation, correcting labels as needed. Regions of uncertainty were marked with asterisks, to be checked by another annotator if necessary. If the second annotator remained uncertain, a third annotator acted as tie-breaker. This approach yielded high inter-rater agreement. Weighted Cohen's kappa with equidistant penalties are 0.87 for stance strength and 0.93 for polarity ($p = 0$), with unweighted kappa for combined labels of 0.88[2].

2.1. Prior analysis

Previous analysis investigated the relationship between labeled stance-related behavior, such as stance strength, and acoustic-prosodic measures such as pitch, intensity, duration, and speaking rate. Pitch and intensity were extracted every 10ms using Kaldipitch [23] and Praat's "To Intensity," respectively. All values were then log-scaled and z-scored normalized on a per-speaker, per-task basis. Spurt-level values were computed, including mean, maximum, and minimum. Duration was extracted from the manual spurt alignment. Speaking rate was computed as words per second, based on the number of words and spurt duration from manual transcription and alignment.

[2]A preliminary analysis employing triple blind annotation yielded kappa of 0.57 for stance strength and 0.69 for polarity.

Significant effects of stance strength were found for all of these measures. Post-hoc showed that, although all levels of stance strength differed significantly in pitch and duration, low and no stance levels did not differ significantly in intensity[19]. Overall, we found that pitch, intensity, duration, and speaking rate all increased as stance strength increased.

3. Energy and Pitch Contour Dynamics

The above measures are well-studied and have been employed extensively in acoustic-prosodic analysis of linguistic behavior at many levels. However, the measures fail to capture the dynamics of the speech signal. This work investigates two measures of speech dynamics, relating to the energy and pitch contour, based on the modulation spectrum and subband autocorrelation measures of pitch change, respectively.

3.1. Modulation Spectrum

The contrasts in speaking rate associated with stance strength and task type [18] can be viewed as energy contours with differing frequency content. This observation suggests an approach where the energy contour is treated as a signal and subjected to Fourier analysis. Shinozaki et al. [24] found that read and conversational speech could be distinguished using a simplified version of the modulation spectrum, via multi-band Fourier analysis of energy contours. A similar measure has been shown to be effective in distinguishing other speaking styles, including adult- and infant-directed speech. These approaches employ a short-term Fourier Transform to extract energy contours and then apply the Fourier transform to the resulting band limited contours.

In our approach, we use a simplified representation with only two energy bands, 0-3kHz and 3-6kHZ. The former focuses on vowel energy and the patterns of syllable rhythm, while the latter emphasized the consonant energy and possible cues to hyperarticulation, as proposed in [24]. Specifically, energy is computed at 100 frames/sec with a 20 ms window. For each band-limited energy contour of a spurt, we find the Fourier Transform of 500 ms overlapping windows and average all such windows in a spurt. Visual inspection of a sequence of spurt energy modulation frequency content suggested that a small number of cepstral features would provide a good low-dimensional representation of the frequency content; 5 coefficients were used in subsequent analyses.

3.2. Pitch Change

Pitch extraction is intrinsically uncertain, and many pitch extraction approaches, such as Praat's pitch trackers, are subject to doubling and halving errors. Viterbi search is often used to enforce a global smoothness constraint over multiple pitch candidates. As an alternative strategy, [25, 26] have proposed a lognormal tied mixture model (LTM). Following the lognormal distribution of pitch values in speech, a tied mixture model is fit with three modes at p, $p/2$, and $2p$, allowing correction or removal of problematic points. Other recent approaches such as [27] have employed machine learning techniques to estimate the likelihood of pitch classes, prior to Viterbi decoding, achieving high performance on the standard Keele database [28].

However, for many tasks such as tone or intonation recognition, it is not the pitch value itself that is of primary interest, but actually the changes in pitch described by the pitch contour. Thus, a robust measure of pitch change may be even more useful in applications that depend on pitch measures.

The subband autocorrelation change detection model (SACD) [29] implemented in the Pitch Change toolbox[3] provides such a model. Similar to [27], this method performs Principal Components Analysis to reduce the dimensionality of the correlogram and applies a multilayer perceptron over this representation to predict pitch classes. However, Pitch Change uses cross-correlation of pitch class probabilities in adjacent frames to compute pitch change values. The use of pitch change measures provided by this technique demonstrated strong effectiveness in Mandarin Chinese tone recognition, both alone and in conjunction with MFCC models [29, 30]. We employ this measure, computed every 10 ms using the Pitch Change toolbox as our base Pitch Change value.

4. Analysis

To assess the utility of dynamic measures of pitch and energy in spoken expressions of stance, we analyze the modulation spectrum based measures of energy (modspec) and subband autocorrelation measures of pitch change (pitchchange) in 11,119 analyzable spurts. For modspec, the measures we employ are the first five cepstral coefficients for each of the two bands, 0-3kHz and 3-6kHz. For each spurt, we compute these values and then z-score normalize them on a per-speaker, per-task basis[4]. For pitchchange, we compute the raw pitchchange values are described above and then compute the maximum, minimum, and mean over each spurt[5].

By ANOVA, we find that there is a significant effect of stance strength on all modspec measures ($p < 0.0001$). Furthermore, post-hoc tests indicate that the third coefficient of the 0-3kHz band differs significantly across all stance strength levels, while all others differ significantly for all but the no-/weak-stance pair. Figure 1 illustrates this contrast. Similarly, by ANOVA, there is a significant effect of stance strength for the pitchchange measures. Post-hoc tests reveal that pitchchange maximum distinguishes all levels of stance strength, while pitchchange mean distinguishes all but the no-/weak-stance pair. Figure 1 demonstrates this contrast in pitchchange by stance strength.

From this analysis, we can see that modspec and pitchchange measures of energy and pitch contour dynamics allows us to identify the full range of distinctions in stance strength levels. Furthermore, we observe an overall trend of increase for both pitch and energy dynamics as stance strength increases. These results echo, and enhance the resolution of, the trends observed for common prosodic measures.

5. Dynamic Measures in Classification

To further investigate the utility of measures of pitch and energy dynamics in analysis of stance-taking behavior, we perform automatic classification of stance strength using these measures. We compare these measures of dynamics to more common prosodic measures and also combine multiple feature types.

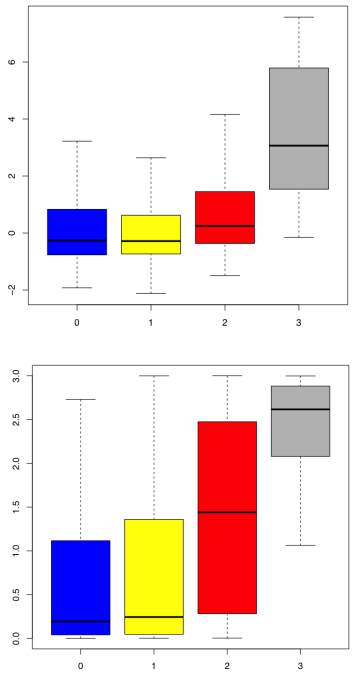

Figure 1: *Above: Modulation spectrum measure z-score and stance strength (0:none, 1:weak, 2:moderate, 3:strong). Below: Pitchchange z-score maximum and stance strength*

Table 1: *Distribution of stance strength classes*

Stance Strength	Proportion
No Stance	27.2%
Weak Stance	49%
Moderate Stance	23.2%
Strong Stance	0.6%

5.1. Classification experiment setting

We perform 4-way stance strength classification, labeling instances as no-, weak-, moderate-, or strong-stance. The distribution of stance strength categories is shown in Table 1. The weak stance class is the most frequent, accounting for approximately 49% of instances in this dataset, followed by relatively similar rates of no- and moderate- stance. The strong stance category is quite infrequent, with fewer than 1% of spurts.

We employ Gradient Boosting Trees (GBT) [31], a generalized boosting method that uses decision trees to perform optimization of arbitrary differentiable loss functions[6]. Proposed for prediction problems with a continuous input space, GBT has proven highly effective in a variety of conversation understanding tasks (e.g.,[32])[7].

Classification is performed in a 10-fold cross-validation setting, splitting the data into 10 subsets and iteratively using 9

[3]Available for download from Microsoft.

[4]Normalization factors are computed based on values between the 10th and 90th percentiles to reduce the effect of outliers.

[5]Pitchchange values are unstable – extremely high or low – at silence boundaries. We heuristically threshold the values to the range -3,3.

[6]We employed the implementation in scikit-learn with the following parameters: n_estimators: 1000, max_leaf_nodes: 4, max_depth: None, random_state: 2, min_samples_split: 5.

[7]We performed exploratory experiments using Support Vector Machines and Deep Neural Networks, but obtained the best effectiveness with GBT when incorporating prosodic features.

subsets for training and the final one for test. We contrast classification using different subsets of classification features, including standard pitch and intensity measures, our new measures of pitch and energy contour dynamics, and text features. We report accuracy as our evaluation metric.

5.2. Classification Features

In these experiments, we focus primarily on contrasting the impact of different classes of prosodic features, and thus perform most of the experiments using prosodic features alone. However, prior work [20] demonstrated that, while prosodic and speaking style features outperformed a most common class baseline on stance strength classification, the accuracy was relatively low, a bit over 50%. In contrast, that work found that word unigrams were the single most effective feature type, yielding 20-40% relative reduction in error, with word and word+punctuation features respectively. Thus, we also include experiments where text-based features, excluding punctuation, are augmented with the new dynamic measures.

For prosodic features, we emphasize the use of the modspec and pitchchange features described above and employed in the acoustic analysis for each spurt. In addition, for contrastive purposes, we employ standard measures of pitch and intensity. Specifically, we extract pitch and intensity every 10 ms using Kaldipitch [23] and Praat's "To Intensity", respectively. These raw values are then log-scaled and z-score normalized on a per-speaker, per-task basis. Based on the manual spurt alignments, we compute per-spurt features of maximum, minimum, mean, median, and standard deviation of both pitch and intensity.

For text features, we extract the manually transcribed text associated with each spurt. Sentences and words are tokenized with NLTK's *sentence_tokenize()* and *word_tokenize()*, respectively. All punctuation is stripped, except for apostrophes in contractions. The tokenized text of the spurt is converted to a binary unigram feature vector the length of the corpus vocabulary, with 1 indicating token presence and 0 indicating absence.

5.3. Results & Discussion

Prosody-only classification We begin by presenting results for prosody-only classification and then turn to a text-based comparison. We find that the new dynamic measures bring important additional information to models using only standard measures of pitch and intensity for stance classification. First we consider individual feature classes in isolation. The standard pitch and intensity measures hover around the common class baseline, at ≈49%. The dynamics-based measures fare slightly better ranging from 49.8% (pitchchange) to 50.8% (modspec).

All measures improve when used in combination, as seen in Table refresultsprosody. Combined pitch and intensity outperform either of the individual measures of pitch and energy contour dynamics. However, combining pitchchange features with standard pitch and intensity provides an additional boost in stance classification accuracy. In addition, combining modspec features with standard pitch and intensity measures also yields a further – and significant – improvement in accuracy over standard pitch and intensity alone, bringing accuracy to 52.3%. Integrating all features yields the best classification accuracy for prosody only stance strength recognition, at 52.7%.

Combination with text-based features Since lexical content plays a key role in stance strength classification, we also assess the effect of combining our different classes of prosodic measures with a text-based binary unigram vector representation. Under the same classifier regime, the binary text vector alone achieves 63.6% accuracy, attesting to the importance of even simple lexical content for this task. Adding prosodic measures of pitch and intensity, alone or in combination, yields little or no improvement. However, the modspec measures of energy contour dynamics, alone or combined with other prosodic features, give some improvement, to 64.2%.

Table 2: *Stance strength classification accuracy under different feature combinations. Feature combination improves effectiveness. P: standard pitch; I: standard intensity; PC: pitchchange; MS: modspec*

Features	Accuracy
PC alone	49.8%
MS alone	50.8%
P+I	51.5%
PC+P+I	52.3%
MS+P+I	52.2%
All features	52.7%

Discussion The classification experiments comparing and combining the new measures of pitch and energy contour dynamics with standard text and prosodic measures indicate that these measures bring additional information to the task of discriminating different levels of stance strength. While the absolute accuracy attests to the difficulty of this task, given inter-speaker variation and a holistic notion of stance strength, it is still interesting to observe that dynamic measures of pitch and energy can improve this conversational understanding task. In addition, the improvements from combining the modspec and pitchchange measures themselves attest to the diversity of cues employed in signaling stance behaviors.

6. Conclusions & Future Work

This work has introduced two novel measures for pitch and energy contour dynamics to the conversational understanding task of recognizing the presence and strengths of attitudes expressed in speech. We investigated the relationship between these new measures and stance strength through both statistical analysis and automatic classification experiments. We found that both measures provided additional discriminating information about the strength of attitude expression beyond standard pitch and intensity measures. Combining all measures for stance strength classification yielded the best results. However, we expect that alternate representations of these and other models of speech dynamics would allow us to better harness prosodic information in the speech signal. Both measures currently aggregate over the unit of analysis, and we expect that finer-grained models of attention and temporal dynamics as well as sequence models of classification would yield further improvements.

7. Acknowledgements

This work has been supported by NSF IIS: #1351034. 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 NSF or the U.S. Government. Thanks also to the ATAROS team, especially Mari Ostendorf, Valerie Freeman, Heather Morrison, and Anna Moroz.

8. References

[1] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, *Longman grammar of spoken and written English*. Longman, 1999.

[2] P. Haddington, "Stance taking in news interviews," *SKY Journal of Linguistics*, vol. 17, pp. 101–142, 2004.

[3] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, *A Comprehensive Grammar of the English Language*. New York: Longman, 1985.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002, pp. 79–86.

[5] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2–3, pp. 165–210, 2005.

[6] S. Somasundaran and J. Wiebe, "Recognizing stances in online debates," in *Proceedings of ACL 2009: Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, 2009.

[7] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of ICASSP-92*, 1992, pp. 517–520.

[8] G. Murray and G. Carenini, "Detecting subjectivity in multiparty speech," in *Proceedings of Interspeech 2009*, 2009, pp. 2007–2010.

[9] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proceedings of Human Language Technologies Conference*, 2001.

[10] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech type," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2002.

[11] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus," in *Proceedings of the Measuring Behavior Symposium on "Annotating and Measuring Meeting Behavior"*, 2005.

[12] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proceedings of HLT-NAACL Conference*, Edmonton, Canada, 2003.

[13] S. Somasundaran, J. Wiebe, P. Hoffmann, and D. Litman, "Manual annotation of opinion categories in meetings," in *ACL Workshop: Frontiers in Linguistically Annotated Corpora(Coling/ACL 2006)*, 2006.

[14] T. Wilson, "Annotating subjective content in meetings," in *Proceedings of the Language Resources and Evaluation Conference*, 2008.

[15] T. Wilson and S. Raaijmakers, "Comparing word, character, and phoneme n-grams for subjective utterance recognition," in *Proceedings of Interspeech 2008*, 2008.

[16] S. Raaijmakers, K. Truong, and T. Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, October 2008, pp. 466–474. [Online]. Available: http://www.aclweb.org/anthology/D08-1049

[17] V. Freeman, "Hyperarticulation as a signal of stance," *Journal of Phonetics*, vol. 45, pp. 1–11, 2014.

[18] V. Freeman, J. Chan, G.-A. Levow, R. Wright, M. Ostendorf, and V. Zayats, "Manipulating stance and involvement using collaborative tasks: An exploratory comparison," in *Proceedeings of Interspeech 2014*, 2014.

[19] V. Freeman, "The phonetics of stance-taking," Ph.D. dissertation, University of Washington, 2015.

[20] G.-A. Levow, V. Freeman, A. Hrynkevich, M. Ostendorf, R. Wright, J. Chan, Y. Luan, and T. Tran, "Recognition of stance strength and polarity in spontaneous speech," in *Proceedings of the IEEE Workshop on Spoken Language Tecnology 2014*, 2014.

[21] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics '08*, 2008.

[22] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9–10, pp. 341–345, 2001.

[23] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of ICASSP*, 2014.

[24] T. Shinozaki, M. Ostendorf, and L. Atlas, "Characteristics of speaking style and implications for speech recognition," *Journal of Acoustical Society of America*, vol. 126, no. 3, pp. 1500–10, 2009.

[25] K. Somnez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody based speaker recognition," in *Eurospeech*, 1996.

[26] X. Lei, M.-H. Siu, M.-Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for mandarin broadcast news speech recognition," in *Interspeech*, 2006.

[27] B. S. Lee and D. Ellis, "Noise robust pitch tracking by subbandautocorrelation classification,," in *Interspeech*, 2012.

[28] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Eurospeech*, 1995, pp. 837–840.

[29] M. Slaney, E. Shriberg, and X. Huang, "Pitch-gesture modeling using subband autocorrelation change detection," in *Proceedings of Interspeech 2013*, 2013.

[30] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, "Highly accurate mandarin tone classification in the absence of pitch information," in *Proceedings of Interspeech 2013*, 2013.

[31] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 2000.

[32] G.-A. Levow and S. Wang, "Employing boosting to compare cues to verbal feedback in multi-lingual dialog," in *Proceedings of IEEE Workshop on Spoken Language Technology 2012*, 2012.