Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Interactional and Pragmatics-Related Prosodic Patterns in Mandarin Dialog

Nigel G. Ward[1,2], *Yuanchao Li*[2], *Tianyu Zhao*[2], *Tatsuya Kawahara*[2]

[1]University of Texas at El Paso, [2]Kyoto University

nigelward@acm.org, {lyc,zhao}@sap.ist.i.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp

Abstract

The roles of prosody vary from language to language. In European languages prosody is largely involved in pragmatics, but this may be less true for other languages, especially tone languages. As a case study this paper examines Mandarin. Using telephone dialog data and a semi-automatic bottom-up analysis method based on Principal Components Analysis, we identify a dozen prosodic patterns in Mandarin which appear to have pragmatic and/or interactional significance. Examination of the overall fraction of prosodic variation explained by different factors also suggests that Mandarin uses prosody heavily for pragmatic functions in dialog.

Index Terms: conversation, spontaneous speech, tone, Chinese, superpositional modeling, unsupervised learning, prosodic constructions, dialog activities

1. Motivation

In European languages prosody is known to convey many pragmatic functions, and prosody-pragmatics mappings have been exploited in many applications, including turn-taking, user modeling, information retrieval, and speech recognition [1, 2, 3, 4, 5]. However tone languages may use prosody less for pragmatic purposes, and to the extent that this is the case, many techniques that use prosody will be less effective.

Whether this is true is an open question. On the one hand, it is often said that the heavy use of pitch for marking tone makes it less available for pragmatic functions, and indeed in tone languages such functions tend to be marked with particles. On the other hand, it is also said that in tone languages "the communicative use of sentence intonation seems to be as free as in nontone languages" [6], and for Mandarin specifically, the topic of this paper, prosodic correlates have been identified for several pragmatic purposes, including focusing, questioning, managing disfluency, marking topic boundaries, expressing emotion, and turn-taking [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. Bottom-up work has also shown that there is much more to prosody in Mandarin than tone: in read speech "tone identities make up only 40-45% of output F_0" [22].

This paper explores two questions:

1. What non-tonal functions does prosody have in Mandarin?
2. In Mandarin, what fraction of prosody is devoted to pragmatic functions, rather than tone?

We investigate using a method that supports the systematic examination of the major prosodic patterns of a language [23].

2. Data and Methods

As dialog is the realm of speech best suited for observing interactional and pragmatic functions, we chose to work with the CallFriend Mandarin corpus of telephone conversations [24]. We used 77 minutes of data: the first 15 minutes or so of files 0928, 4249, 4257, 4198, and 5975.

2.1. A Superpositional Model

We use a superpositional model, in which the observed prosody is assumed to be the result of summing the effects of simultaneously active underlying factors. Superpositional models have been used for many purposes, for many languages, as surveyed elsewhere [25, 23], including Mandarin [26, 27, 28, 29]. While superposition alone is clearly not adequate for modeling all aspects of prosody, a superpositional model can be a useful approximation.

Among the many possible ways to build a superpositional model, we chose Principal Components Analysis (PCA) because it has been useful in identifying pragmatics-related prosodic patterns in other languages [30]. Our strategy was to compute numerous low-level features at each point in the recordings, and then use PCA to discover the underlying patterns [25, 23]. Specifically, we computed features at 874,000 timepoints t, sampled evenly every 10 milliseconds throughout the dialogs, considering both tracks. At each timepoint these features characterized the activity in that vicinity, and all this data was fed to PCA to discover the underlying factors. Although PCA has some quirks, including a tendency to output factors with symmetric feature loadings, it has the advantage of simplicity.

This assumption-free strategy can be contrasted with that of [22] who started with a model of the "lower-level effects" of tone and then examined the residue when these were eliminated. We choose an assumption-free strategy because, despite many advances in the modeling of tone [31, 32], we are working with noisy and complex dialog data, where building a reasonable baseline model of the "lower-level effects" of tone would be an extreme challenge.

2.2. A Large Set of Time-Spread Features

To broadly characterize the pattern of activity in the vicinity of any point in time, t, we use features computed over windows that together span from about 3 seconds before t to 3 seconds after. The window sizes are roughly proportional to the distance from t, thus, for example, the most distant intensity window is 1.6 seconds long, from –3.2 s to –1.6 s, and the closest is 50 milliseconds long, from –50 ms to 0 ms. There are windows for features of both speakers, enabling the discovery of joint behaviors comprising contributions by both.

Windows are fixed in offset from t, rather than being aligned to a turn, phrase, utterance, word, or syllable. This is so that they can be everywhere-computable and robust, as is necessary to enable the discovery of joint patterns of prosodic behavior by both speakers, since the units of the two speakers are not generally aligned. We also do no time-warping or other stretch-

 doi: 10.21437/SpeechProsody.2016-253

	number of features		
	left track	right track	total
high pitch	20	20	40
low pitch	20	20	40
narrow pitch	10	10	20
wide pitch	10	10	20
intensity	16	16	32
creakiness	14	14	28
rate	10	10	20
Total	100	100	200

Table 1: Prosodic Feature Counts

ing. Thus these features are less well-suited for accurately capturing syllable-level phenomena than they are for other aspects of prosody, including wider-scope and joint-behavior patterns.

While most previous work on Mandarin has focused on intonation, other aspects of prosody are also significant. Accordingly we include features to cover the four commonly-used aspects of prosody — intensity, pitch range, pitch height, and speaking rate — plus creakiness. More details of the computation are given elsewhere [33], with the full description at [34], where the code itself is also freely available. We note, however, that for speaking-rate we use a proxy based on frame-to-frame energy variation, as this has worked well for dialog before [4], and that we use robust, everywhere-valued pitch features to represent pitch height and pitch range [34]. The features were chosen as a set adequate to cover the prosodic features involved in pragmatic and other functions in a number of languages [30], and then, since Mandarin is known to have rapid pitch movement, augmented with 32 extra pitch-height features to obtain finer temporal resolution. Pitch slope features are not explicitly calculated, but, as seen below, patterns in pitch-level changes are revealed by the PCA. Table 1 lists the numbers of features of each type.

Wanting the top factors to be important ones, rather than those which merely explained the most raw variation, before applying PCA we z-normalized all features. Nevertheless the larger number of pitch features biases factors which relate heavily to pitch to come out nearer the top.

Applying PCA to these features gives a dimensional model of Mandarin prosody, in which the observed prosody at each timepoint in the data is explained as the result of multiple simultaneously-active factors, that is, the dimensions.

2.3. Eclectic Interpretation of Patterns

Given the factors (dimensions) output by PCA, our next step was to distinguish those related to tone and those related to pragmatic functions. To do this we used a qualitative inductive method. At each timepoint in the data each factor has a value, which can be positive, negative, or zero. Most informative are the timepoints where a factor has an extreme value. For example, listening to timepoints where Dimension 4 had a high value, most of them obviously involved a backchannel by the speaker in the left track, frequently *oh* 哦 . We further listened to seek pragmatic or semantic commonalities among the instances high or low on a factor. For example, the Dimension 4 positive examples frequently involved a speaker giving new information to the other, the listener demonstrating that they had received this new information, and the speaker then continuing on the same topic.

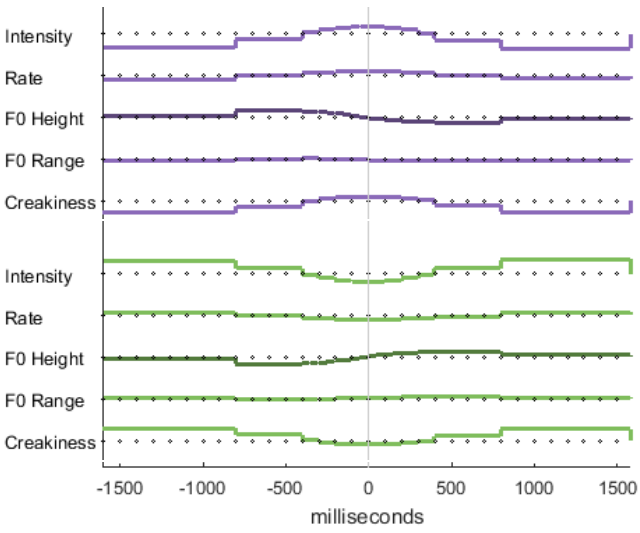

Figure 1: Loadings of Dimension 4. The top of the diagram (purple lines) shows the loadings for the left-track speaker, across the various features and windows, and the bottom half (green) the right-track speaker. Time is in milliseconds. The dotted lines are zeros, with points above them positively loaded and points below negatively loaded.

Conversely, at timepoints where Dimension 4's value was very negative there was generally a backchannel in the right track. Thus for this, and all dimensions, we were able to identify two patterns, one when the value on the dimension was positive and one when negative.

We also examined the loadings. For example, the features loading positively on Dimension 4 include the left-track speaker's intensity over a short interval, as seen in the first line of Figure 1; this of course corresponds to the backchannel. The figure also shows a tendency for the right speaker's pitch to go low about 300-400 milliseconds before the backchannel onset. (It is important to stress that the y-axes in these graphs are not based directly on the features; thus the line labeled "pitch height" is not directly the average or typical pitch height, but rather represents the difference between the loadings of the high-pitch and low-pitch features. Similarly the "pitch range" line shows the difference between the wide-pitch and narrow-pitch features. This is because simple pitch-height features are not robust enough, as noted earlier, and because of the use of PCA. Nevertheless human pitch is generally continuous, and in practice these lines always do indicate pitch contours that occur frequently in the data.)

Interpreting the patterns was time-consuming. For each we listened repeatedly to about a dozen examples of each, and considered the words said, their apparent pragmatic intention and effect, the larger context, and, of course, the tones. For some patterns we were able to find commonalities with few exceptions, but for others we were able to identify only tendencies or families of related functions. This is to be expected: in a superpositional model, the actual meaning is the sum of the meanings of all patterns simultaneously present, and at any specific timepoint the meaning of one pattern can obscure the contributions of another. Nevertheless, each of the descriptions below was valid for at least two thirds of the examples examined.

3. Prosodic Pattern and Interpretations

This section concisely describes some of the patterns observed. The complete loadings for each factor are at http://www.cs.utep.edu/nigel/mandarin/.

Factor 1 related to who was mostly talking.

1 positive **talking** The left-track person is silent while right-track person talks.

1 negative **silent** Conversely, the left-track person silent while the right-track person talks.

Factor 2 related to turn-taking, and was most strongly present, either positively or negatively, when one speaker stopped speaking and the other immediately started.

2 pos **turn take** Taking $t = 0$ as the middle of the turn transition, this pattern prototypically includes:
–1500 ms right-track person increases loudness, speaking rate, and creakiness
–800 ms right-track person begins to lower pitch
–400 ms right-track person's loudness and pitch drop
+400 ms left-track person takes the turn, starting with moderately high pitch and creaky voice
+800ms left-track person's volume and pitch height abate, and the speaking rate increases

2 neg **turn yield** The converse pattern, where left-track person yields the turn and/or the right-track person takes the turn.

Factor 3 had, for both speakers, the same weightings for all features.

3p **negative evaluation** The speaker reports something that someone else has done or felt and assesses it as negative in some way. Within a region of relatively high pitch and fast speaking rate, the speaker shifts to use even higher pitch and increased volume and pitch range over about 1.5 seconds.

3n **brief pause** The speaker pauses for a second or so, while thinking how to express something.

Factor 4 reflected a joint behavior, as mentioned earlier.

4p **new-information backchannel** As seen in Figure 1, this pattern has several specific properties. Taking $t = 0$ as the center of the backchannel, and roughly noting the offset times seen, these include:
–800 ms: a region of increased volume and creakiness for a second or less, generally including the point of new information
–600 ms: a region of low pitch for a half second or less
–300ms: the start of the backchannel, starting moderately high and then falling in pitch, often some form of *oh* 哦
+300ms: the end of the backchannel
+500ms the original speaker's resumed speech, starting with high volume and creakiness

4n **new information and backchannel cue** The converse pattern.

Factor 5 had loadings for both speakers the same.

5p **no new information** One or both speakers was speaking quickly, with lowish pitch, but providing no new information, often saying something that was already clear from the context. One speaker sometimes backchanneled, but, unlike those in Factor 4, these backchannels were quiet, short or fast, and low in pitch, for example *aa-aa* 啊 , and were not aligned with the other speaker's utterances.

5n **topic progression**, One speaker introduced a new facet of the topic, for example progressing from talking about the other speaker's interview to their own interview, or from talking about one person's possible move to start a new job to the question of how her health condition could constrain the move. This pattern involved an overall slow speaking rate and a second or two of high pitch.

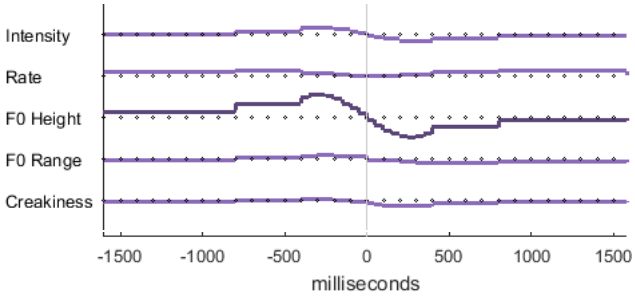

Figure 2: Dimension 10, Loadings negated to show the negative-side pattern.

Factor 6 had opposite loadings for the two speakers.

6p **high activity** The left-track speaker continues speaking or starts or restarts a sentence, speaking quickly and with a short slightly higher pitch, while the right-track speaker is passive, at most producing a long backchannel or filler.

6n **low activity** The converse pattern, with the speaker roles reversed.

Factor 7 was prototypically a joint pattern, where one speaker bid for empathy and the other gave it.

7p **empathy bid** The right-track speaker bids for empathy, either showing exasperation or displeasure (signalled for example with a sigh like *ai-ya* 哎呀), or reporting being in a difficult situation, or inviting the other or asking a question and hoping for a favorable response. The pitch is high over a few seconds, and there is a sudden drop in loudness.

7n **giving empathy** The right-track speaker typically says something sympathetic or agreeable, with low pitch.

Factor 8 had loadings for both speakers the same.

8p **engagement** One or both speakers are engaged in a topic. Prosodically there is a second or less with moderate and falling volume, slight creakiness, and ending with a lower pitch. This sometimes involves modal particles such as *ma* 嘛 , *ba* 吧 , and *a* 啊 .

8n **topic exhaustion** Both speakers are ready to end the current topic, and there is a short pause in the speech or a low-volume fast mumble.

Factor 9 had opposite loadings for the two speakers.

9p **cue for response** The left-side speaker asks the other a question, makes a suggestion, or otherwise invites or cues a short response. This involves a region of high creakiness and loudness, followed by a word with high pitch and increasing volume. The right-side speaker briefly answers, accepts, agrees or produces a backchannel to show understanding.

9n **cued response** The converse pattern.

Factor 10, with loadings for both speakers the same, comprises two apparently unrelated patterns.

10p **contrast** This involves a contrast, for example between some past situation and now, between one speaker's location or situation and the other's, or between an inference made by the listener and reality. This involves a sharp rise

in pitch over about 800 milliseconds with a slight increase in creakiness, for example over the word *fǎn zhèng* 反正 (*anyway*).

10n **falling tone** As seen in Figure 2, this involves a salient long pitch drop, and appears to involve emphasis of a word containing tone 4. There is also a slight decrease in volume and creakiness.

To mention briefly the apparent roles of two more dimensions: Factor 11 seems to involve detachment, where either the right or the left speaker is contributing only in a perfunctory way, without much warmth or interest. Factor 12 seems to involve some form of divergence, such as a difference of opinion or knowledge asymmetry between the speakers.

4. The Magnitude of Pragmatics-Related Prosody

Turning now to our second question, regarding the relative importance of the different functions of prosody, we first must acknowledge that any attempt to quantify this is rife with difficulties. Nevertheless, a very rough estimate can be obtained by considering the total amount of variation explained by factors with clear pragmatic functions, as summarized in Table 2. The sum of the variation in dimensions 2–12, excluding the half of dimension 10 that relates to tone, is 0.30. The total effective amount of variation, is not 1.0 but 0.77, excluding dimension 1, which relates mostly to the brute fact of which person is speaking. The former is 39% of the latter, and thus we estimate the pragmatic load of prosody as at least 39% of the total. For an even rougher estimate of the relative load of tone, we scanned the loadings of other dimensions and found that most factors below number 16 relate mostly to very localized pitch-height features; thus most of the remaining prosodic variation may be devoted to tone.

This is compatible with Tseng and Su's finding that 55-60% of the variation is not explicable by individual tones alone, although the two results are not quite comparable, since we used dialog rather than read speech, we used unaligned features, and we included features beyond pitch height. While any such estimate must be regarded with skepticism, as it is dependent on the features chosen, these estimates do suggest that pragmatic functions are a major part of prosody's role in Mandarin dialog.

5. Discussion and Summary

Interestingly, many of these pragmatic functions are also conveyed with prosodic patterns in other languages, and for at least two of the functions, empathy bids and backchannel cues, some of the prosodic components are the same also in English, Spanish, and Japanese. Investigation of possible universals is a prime topic for further research.

Another topic for future work would be to investigate these patterns using other methods. While our method, statistical pattern discovery plus subjective interpretation, is good for discovering new patterns and their possible significance, it is not adequate to establish either. Future work might reexamine the observed tendencies experimentally. Among other things, this might help pin down the exact timing and other properties of the components of these patterns, as our method gives only a blurred picture.

Nevertheless, our exploration suggests that Mandarin uses prosody for many pragmatic functions. Some details of the patterns found corroborate observations in the literature, for example regarding back-channeling, turn yielding, and turn starts [11, 9, 20]. Other patterns appear to convey functions not previously related to prosody in Mandarin, including negative evaluation, bids for empathy, topic progression, topic exhaustion, and contrast. We conclude that a significant fraction of the prosody of Mandarin dialog relates to pragmatic functions, and that developers interesting in using prosody for various applications should not shy away from trying their methods for Mandarin.

	var.	side	type	interpretation
1	23%	pos	neither	speech
		neg	neither	silence
2	5%	pos	prag	turn take
		neg	prag	turn yield
3	4%	pos	prag	negative evaluation
		neg	prag	brief pause
4	3%	pos	prag	new-info. backchannel
		neg	prag	backchannel cue
5	3%	pos	prag	no new information
		neg	prag	topic progression
6	3%	pos	prag	high activity
		neg	prag	low activity
7	2%	pos	prag	empathy bid
		neg	prag	showing empathy
8	2%	pos	both	engagement
		neg	prag	topic exhaustion
9	2%	pos	prag	short answer or response
		neg	prag	question or other response cue
10	2%	pos	prag	contrast
		neg	tone	falling tone

Table 2: Summary of Findings for the Top Ten Dimensions (Top 20 Patterns). Var. is the variance explained by each factor.

6. Acknowledgments

This work was supported in part by a Fulbright Award and by DARPA under the Lorelei program. This work does not necessarily reflect the position of the Government, and no official endorsement should be inferred. We thank Wen Wang for comments.

7. References

[1] J. Gratch, N. Wang, A. Okhmatovskaia, F. Lamothe, M. Morales, R. van der Werf, and L.-P. Morency, "Can virtual humans be more engaging than real ones?" *Lecture Notes in Computer Science*, vol. 4552, pp. 286–297, 2007.

[2] K. Forbes-Riley and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," *Speech Communication*, vol. 53, pp. 1115–1136, 2011.

[3] N. G. Ward, S. D. Werner, F. Garcia, and E. Sanchis, "A prosody-based vector-space model of dialog activity for information retrieval," *Speech Communication*, vol. 68, pp. 86–96, 2015.

[4] 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.

[5] S. R. Gangireddy, S. Renals, Y. Nankaku, and A. Lee, "Prosodically-enhanced recurrent neural network language models," in *Interspeech*, 2015.

[6] A. S. Abramson and K. Svastikula, "Intersections of tone and intonation in Thai," in *Status Report on Speech Research, Haskins Laboratories*, 1983, pp. 143–167.

[7] J. Yuan and D. Jurafsky, "Detection of questions in Chinese conversational speech," in *Automatic Speech Recognition and Understanding, IEEE*, 2005, pp. 47–52.

[8] G.-A. Levow, "Assessing prosodic and text features for segmentation of Mandarin broadcast news," in *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*, 2004, pp. 28–32.

[9] ——, "Turn-taking in Mandarin dialogue: Interactions of tones and intonation," in *Proc. SIGHAN Workshop*, 2005, pp. 72–78.

[10] F. Lie, Y. Xu, S. Prom-on, and A. C. L. Yu, "Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modeling," *Journal of Speech Sciences*, vol. 3, pp. 85–140, 2013.

[11] N. G. Ward and J. L. McCartney, "Visualizations supporting the discovery of prosodic contours related to turn-taking," in *Interdisciplinary Workshop on Feedback Behaviors in Dialog*, 2012.

[12] X.-L. Zeng, P. Martin, and G. Boulakia, "Tones and intonation in declarative and interrogative sentences in Mandarin," in *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, 2004.

[13] M. Grubic, "The prosody realization of y/n-questions in Mandarin Chinese," 2008, University of Potsdam, Linguistics Institute, Master's Thesis.

[14] C.-K. Lin and L.-S. Lee, "Improved features and models for detecting edit disfluencies in transcribing spontaneous Mandarin speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, pp. 1263–1278, 2009.

[15] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1145–1154, 2006.

[16] A. Li, *Encoding and Decoding of Emotional Speech: A Cross-Cultural and Multimodal Study between Chinese and Japanese*. Springer, 2015.

[17] S.-C. Tseng, "Repairs in Mandarin conversation," *Journal of Chinese Linguistics*, vol. 34, pp. 80–120, 2006.

[18] R.-J. R. Wu, "Managing turn entry: The design of EI-prefaced turns in Mandarin conversation," *Journal of Pragmatics*, vol. 66, pp. 139–161, 2014.

[19] W. Zhang, "A prosodic analysis of insertion repair at transition space in Chinese conversation," in *International Conference on Asian Language Processing*, 2014, pp. 151–153.

[20] Y.-F. Liu and S.-C. Tseng, "Linguistic patterns detected through a prosodic segmentation in spontaneous Taiwan Mandarin speech," in *Linguistic Patterns in Spontaneous Speech*. Institute of Linguistics, Academia Sinica, 2009, pp. 147–166.

[21] K. Schack, "Comparison of intonation patterns in Mandarin and English for a particular speaker," *University of Rochester Working Papers in the Language Sciences*, vol. Spring, pp. 24–55, 2000.

[22] C.-Y. Tseng and Z.-Y. Su, "What's in the F0 of Mandarin speech: Tones, intonation and beyond," in *6th International Symposium on Chinese Spoken Language Processing*. IEEE, 2008, pp. 1–4.

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

[24] A. Canavan and G. Zipperlen, *CALLFRIEND Mandarin Chinese Speech*. Linguistic Data Consortium, 1996, lDC Catalog No. LDC96S55, ISBN: 1-58563-070-5.

[25] 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.

[26] G.-P. Chen, G. Bailly, Q.-F. Liu, and R.-H. Wang, "A superposed prosodic model for Chinese text-to-speech synthesis," in *International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 177–180.

[27] C.-Y. Tseng, S.-H. Pin, Y. Lee, H.-M. Wang, and Y.-C. Chen, "Fluent speech prosody: Framework and modeling," *Speech Communication*, vol. 46, pp. 284–309, 2005.

[28] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, "Unsupervised joint prosody labeling and modeling for Mandarin speech," *The Journal of the Acoustical Society of America*, vol. 125, pp. 1164–1183, 2009.

[29] Y. Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, pp. 85–115, 2011.

[30] N. G. Ward and P. Gallardo, "Non-native differences in prosodic construction use," 2015, submitted.

[31] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, "A statistics-based pitch contour model for Mandarin speech," *The Journal of the Acoustical Society of America*, vol. 117, pp. 908–925, 2005.

[32] C. Shih and H.-Y. D. Lu, "Effects of talker-to-listener distance on tone," *Journal of Phonetics*, vol. 51, pp. 6–35, 2015.

[33] N. G. Ward and S. Abu, "Action-coordinating prosdy," in *Speech Prosody*, 2016.

[34] N. G. Ward, "Midlevel prosodic features toolkit," 2015, http://www.cs.utep.edu/nigel/midlevel/, https://github.com/nigelgward/midlevel.