Visualizations Supporting the Discovery of Prosodic Contours Related to Turn-Taking

Nigel G. Ward, Joshua L. McCartney

Department of Computer Science, University of Texas at El Paso, USA

nigelward@acm.org, jlmc@miners.utep.edu

Abstract

Some meaningful prosodic patterns can be usefully represented with pitch contours, however developing such descriptions is a labor-intensive process. To assist in the discovery of contour representations, visualization tools may be helpful. Edlund *et al.* [1] proposed the superimposition of hundreds of pitch curves from a corpus to reveal the general patterns. In this paper we refine and extend this method, and illustrate its utility in the discovery of a prosodic cue for back-channels in Chinese.

Index Terms: prosodic cue, tune, turn-taking, back-channel, Chinese, bitmap cluster, overlay, superimpose

1. Why Contours?

In human dialog, turn-taking is largely managed by means of prosodic signals, or cues, exchanged by the participants. A dialog system that can correctly recognize and respond to these cues may be able to make the user experience more efficient and more comfortable [2, 3, 4]. These cues often seem to involve pitch contours, or tunes: specific patterns of ups and downs over time. Figure 1 shows three examples, diagrammed in various styles.

However, in the spoken dialog systems community, those working on turn-taking generally do not use contours, neither explicitly nor implicitly. Rather the "direct approach" [5] has become mainstream. In this method, numerous low-level prosodic features are computed and fed into a classifier trained on the decision of interest, for example, whether to initiate a turn or wait. This method has seen many successes.

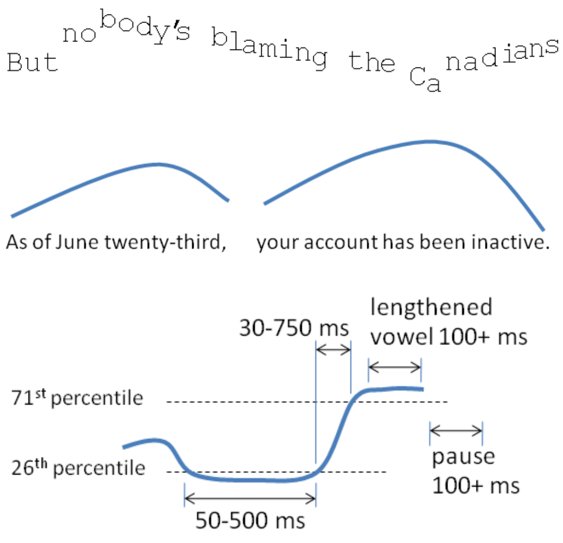

Figure 1: Examples of pitch contours: a contradiction contour (after [6], pg 246), a nonfinal contour followed by a final contour (after [7], pg 183), and a back-channel cuing contour for Spanish (from [8], building on [9], see also Figure 6).

However contours also have their merits. A description in terms of a contour can be concise and may possess more explanatory power than a complex classifier. A contour-based description may apply more generally to other dialog types and other domains of discourse, whereas a complex classifier may perform well only for the corpus it was trained on. In some ways a contour may be a more natural description of a prosodic pattern. For one thing, describing a pattern in terms of low-level features presents some choices which may lack real significance: for example the two descriptions "pitch rise" and "low pitch followed by a high pitch" refer to different mid-level features that may not actually differ in realization; however if drawn as contours their similarity is obvious. As another example, when describing a pattern in terms of features the temporal dependencies may be obscure, as in a rule which requires "low at t – 700" and "high at t – 400", but with contours, the sequencing and timing of the components is immediately clear.

Another advantage of contours is that people who need to know the effective prosodic patterns of a language, for example second language learners, can understand such diagrams fairly quickly. It is even conceivable that contours approximate the true nature of these prosodic patterns as they exist in the human mind. The notion of cue strength [10, 11] may have a natural implementation in terms of contours: the similarity between an input pitch curve and the cue contour may be an easy way to estimate cue strength. Contour-based descriptions can be used not only for recognition but also for production. And finally, contours and the parameters describing them may serve as useful higher-level features for classifiers.

2. The Difficulty of Discovering Contours

Despite the attractions, contours have one great disadvantage: the difficulty of discovering them. In contrast to the direct method, where, as long as one has the necessary resources and properly prepared data, the hard work can be entrusted to the machine learning algorithm, the discovery of a new prosodic contour can be a time-consuming process. Of course any specific utterance has a pitch contour, but going from examples to a general rule is not straightforward.

In particular, elicitation and instrumental techniques that work for monolog phenomena are hard to apply to dialog-specific phenomena such as the prosody of attitude, information-state, speaker and interlocutor cognitive state, and turn-taking. While some tools and methods are designed to support the discovery of dialog-relevant prosodic patterns [12, 13], the process is generally still labor-intensive.

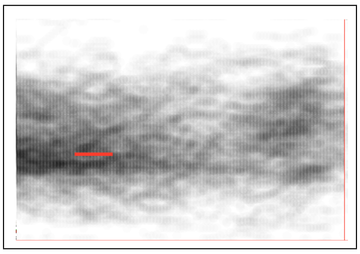

Figure 2: from Figure 6 of Edlund *et al.* [1], by permission

In 2009, however, Edlund, Heldner and Pelcé proposed the use of “bitmap clusters” [1], a visualization method based on superimposing many individual pitch contours to reveal general patterns:

> by plotting the contours with partially transparent dots, the visualizations give an indication of the distribution of different patterns, with darker bands for concentrations of patterns

They used this method to visualize the contexts preceding various utterance types in the Swedish Map Task Corpus. Figure 2, from their paper, shows the contexts preceding 859 talkspurts which were tagged as “very short utterances” and as having “low propositional content”, which were probably mostly acknowledgments. The possibly visible red rectangle was added by hand to mark the frequent occurrence of a region of low pitch found “860 to 700ms prior to the talkspurts”, which they identified with a back-channel cue previously noted in the literature.

Bitmap clusters are, however, still only suggestive, and so far have not been shown useful for new discoveries. This paper builds on foundation to create visualization methods that are.

3. Visualization Improvements

We made several improvements to [1].

First, we chose the pitch regions to overlap in a different way. Edlund *et al.* aligned the ends of the talkspurts at the right edge of the display, presumably assuming that the prosodic cues of interest occur at, and are aligned with, utterance ends. While possibly valid for some dialog types, this is not suitable for, say, back-channels in dialog, which often overlap the continuing speech of the interlocutor. We therefore aligned based on the start of the response of interest, as was also done by [14] for speaking fraction and gaze. Thus our right edge, the 0 point, is always the onset of the response.

Second, we normalized the pitch differently. Edlund *et al.* vertically aligned the contours so that the “median of the first three voice frames in each contour fell on the midpoint of the y-axis,” providing a form of per-utterance normalization. We chose instead to normalize per-speaker, based on our experience that normalization with respect to longer time spans can improve identification of cues [15], probably because turn-taking signals, unlike some other prosodic phenomena, are not tightly bound to utterances, but are relative to the speaker’s overall behavior. Among the various possible normalization schemes, we chose a non-parametric approach, representing each pitch point as a percentile of the overall distribution for that speaker. Compared to approaches which explicitly estimate parameters, such as pitch range or standard deviation, using assumptions about the distribution, we felt this likely to be more robust.

Third, we chose to display an additional feature, energy, again normalized by speaker and expressed in percentiles. This was for two reasons. First, the pattern of speaking versus silence is also important for turn-taking, and we wanted to represent and model this explicitly, rather than leave it to some generic utterance-delimiting pre-processing phase. Second, energy is important in identifying stressed syllables, fillers and so on.

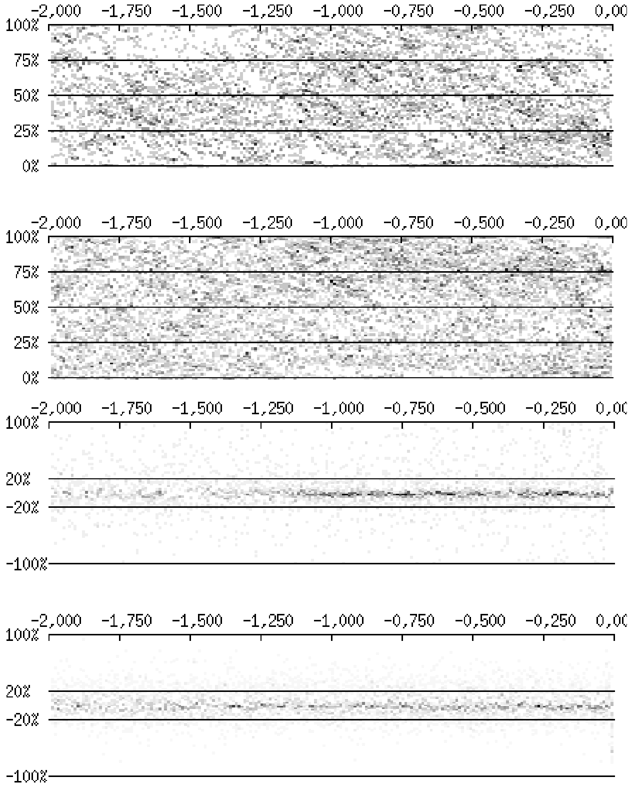

Figure 3: Overlaid Pitch, Energy, Delta Pitch and Delta Energy for Japanese

Fourth, we included the deltas: delta pitch and delta energy. Delta pitch may reveal upslopes, downslopes, and flat regions, and delta energy may reveal lengthened syllables and slow speaking rate. For pitch, if two adjacent pitch points are valid we plot the difference between the previous pitch value and the current pitch value, both expressed as percentiles.

Fifth, we extended the displays out to 2 seconds of past context, to look for longer-term patterns.

Sixth, we did without pitch smoothing, not wanting to risk losing information.

Seventh, since what we really want to see is not the distributions before the events of interest, but how those distributions differ from the general distributions, we added a sharpening step. This was done by subtracted out the global average distribution from each offset. In other words, we subtracted the mean for each percentile, using means estimated from a fairly large random sample over the dialogs. Before doing this the diagrams were blurry and hard to interpret; afterwards they were much sharper, although somewhat more blotchy.

Henceforth we will refer to diagrams made in this way as “overlaid prosodic displays”. Each point represents the count of times that value occurred at that time, normalized so that the highest-count point is pure black.

4. Initial Validation

We developed these refinements as we tried to better visualize to the contexts preceding back-channels in several languages.

We chose to look at back-channeling because it is a classic issue in turn-taking, and because previous research suggests that, among all turn-taking phenomena, back-channeling may be the one where the behavior of one speaker is most strongly influenced by the immediately preceding prosody of the other. The result for Japanese is seen in Figure 3, showing the overlaid prosodic displays for the speech of the interlocutor in the contexts immediately preceding 873 back-channels in casual conversation [15].

While no contour is directly visible, some useful features are: in the second or so preceding the back-channel, the interlocutor's pitch tends to be low, around the 25th percentile, starting around 200ms before the back-channel, and stable; and the energy tends to be high starting about 1000ms before the back-channel, but never very loud in the final 200ms. This roughly matches what we know: that the primary cue to back-channels in Japanese is a region of low pitch, with the interlocutor usually continuing speaking at least until the back-channel response starts. While the optimal prediction rule we found earlier looks somewhat different (requiring a region of pitch lower than the 26th percentile 460 to 350 ms before the back-channel onset [15]) this visualization could clearly be a useful clue to the discovery of such a rule. Applying this method to English, Egyptian Arabic, and Iraqi Arabic data also revealed patterns which matched what we know from previous work [8].

5. Utility for Cue Discovery

Of course, interpreting a diagram is easy when you already know what you expect to see. As a fairer test of the utility of this visualization, we applied it to a language which we had not previously examined, Chinese.

Using 18 dialogs from the Callhome corpus of telephone speech [16], 90 minutes in total, we had two native speakers independently identify all back-channels according to the criteria of [15]. One identified 528 and the other 467. We then took the intersection of the two sets, reasoning that working with unambiguous cases would make it easier to see the normal pattern. This gave us 404 back-channels.

Digressing briefly to comment on back-channeling in Chinese, contrary to what is sometimes reported, back-channels were quite frequent: at over 4 per minute, almost as common as in English. This however may be due in part to the fact that at least one participant in each dialog was resident in North America. Also, although not important for current purposes, we had the annotators label the back-channels. As they were not phonetically sophisticated, we let them use whatever letter sequences they liked. The fifteen most frequent labels of one annotator were *uh*, *oh*, *dui*, *uh-huh*, *em*, *shima*, *hmmm*, *ok*, *yeah*, *huh*, *duia*, *uhuh*, *shia*, *hmmmm*, and *good*, similar to those seen in other corpora [17].

The task we set ourselves was that of discovering what prosodic pattern in the interlocutor's speech was serving to cue back-channel responses. We formalized this in a standard way [15, 2, 18], requiring a predictor able to process the dialog incrementally and, every 10 milliseconds, predict whether or not a back-channel would occur in the next instant, based on information in the interlocutor's track so far. The second author, armed with the visualizations seen in Figure 4 and software infrastructure previously developed for extracting prosodic features and making similar decisions for other languages, but with no knowledge of Chinese, got to work.

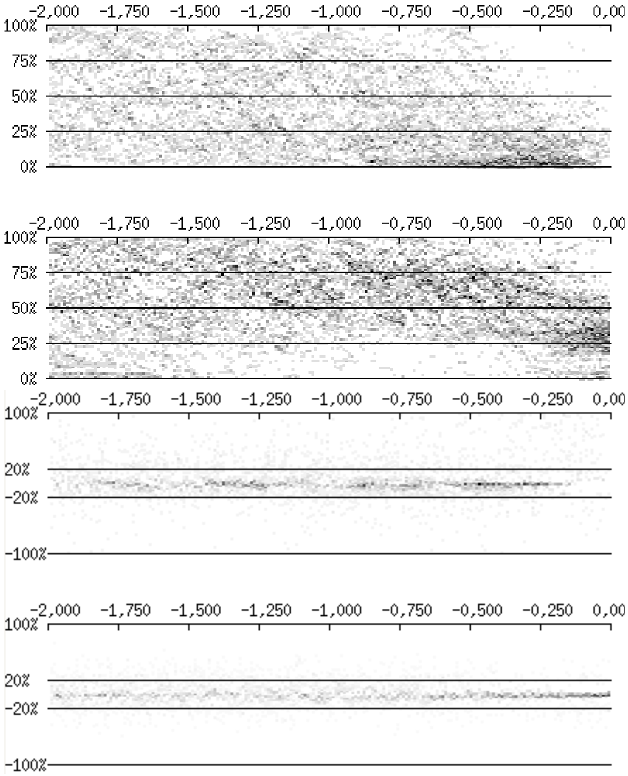

Figure 4: Overlaid Pitch, Energy, Delta Pitch and Delta Energy for Chinese

He immediately noted that the pitch tends to go extremely low from about –500 to –100 milliseconds, and that the energy went low starting at about –200 milliseconds, although not necessarily to silence. The deltas indicated that the pitch tended to be flat from –500 to –200ms, and that the energy also tended to be stable from –600 to 0. Before long he came up with a predictive rule: in Chinese, respond with a back-channel if the interlocutor's speech contains:

- a low pitch region, below 15% and lasting at least 220ms, followed by
- a pause of at least 150ms

This predicted back-channel occurrences with 25% coverage and 9% accuracy. Improvement is certainly possible, but the performance is well above random guessing, which gives 15% coverage and 2% accuracy.

6. Future Work

Thus we conclude that overlaid prosodic displays are a visualization method with value. However it clearly has room for improvement. Consider Figure 5, displaying the contexts of 152 back-channels in Spanish [9]. While it suggests some features likely to be components of a contour, including some fuzzy pitch tendencies, a pause in the last quarter second, and possibly a tendency for flat pitch from –1500 to –1000 ms, as seen in the deltas, there is not much else, even knowing the pattern we expect to find (Figures 1 and 6).

There are several possibilities for improving these visualizations. One could explicitly display also duration or rate. One could make the features more robust. One might apply a thinning algorithm to visually accentuate the tendencies: to turn cloudy streaks into nice curves. One could improve the way the horizontal alignment is done in generating the overlays. In particular, as reaction times vary, the time from the prosodic cue,

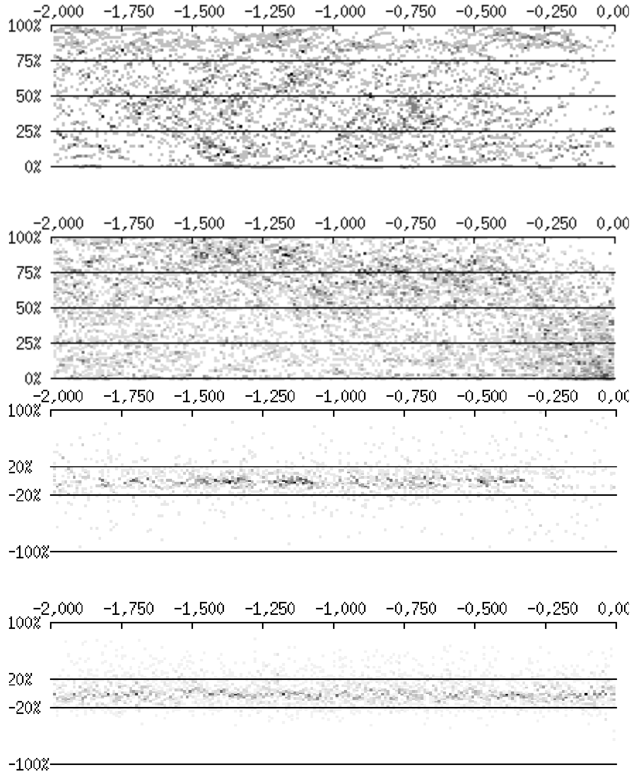

Figure 5: Overlaid Pitch, Energy, Delta Pitch and Delta Energy for Spanish

whatever it may be, to the response will not be constant, so one might devise an expectation-maximization algorithm, where the horizontal alignments are iteratively adjusted to make the pitch contours align better. Finally one could use parametric methods [19] to force the visualizations to really look like contours.

We do not propose contours as a panacea. Among other weaknesses, they do not naturally represent the degree to which their component features may stretch internally or relative to each other; for that purpose they need annotations (as in Figure 1) or a sibling description (as in Figure 6). However contour-based descriptions can be useful, and this new tool can help with their discovery.

Predict a back-channel starting 350 ms after all of the following conditions hold:

A. a low-pitch region,
    < 26th percentile and 50–500ms in length
B. a high-pitch region for at least one moment,
    starting > 75th percentile and never < the 26th
C. a lengthened vowel of duration >= 100 ms
D. a pause >= 100ms

Where:
B closely follows A:
    30-750 ms from end of low to start of high
C closely follow B:
    0-100ms from start of high to start of lengthened vowel
D closely follows C:
    0-60ms from end of lengthened vowel to start of pause

Figure 6: A rule for predicting back-channel opportunities in Spanish, from [8]

Acknowledgment: This work was supported in part by the NSF as Project No. 0415150 and by RDECOM via USC ICT. We thank Tatsuya Kawahara for comments.

7. References

[1] J. Edlund, M. Heldner, and A. Pelcé, "Prosodic features of very short utterances in dialogue," in *Nordic Prosody - Proceedings of the Xth Conference*, pp. 56–68, 2009.

[2] 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.

[3] A. Raux and M. Eskenazi, "A finite-state turn-taking model for spoken dialog systems," in *NAACL HLT*, 2009.

[4] G. Skantze and D. Schlangen, "Incremental dialogue processing in a micro-domain," in *EACL*, pp. 745–753, 2009.

[5] E. E. Shriberg and A. Stolcke, "Direct modeling of prosody: An overview of applications in automatic speech processing," in *Proceedings of the International Conference on Speech Prosody*, pp. 575–582, 2004.

[6] D. Bolinger, *Intonation and Its Parts*. Stanford University Press, 1986.

[7] M. H. Cohen, J. P. Giangola, and J. Balogh, *Voice User Interface Design*. Addison-Wesley, 2004.

[8] N. G. Ward and J. L. McCartney, "Visualization to support the discovery of prosodic contours related to turn-taking," Tech. Rep. UTEP-CS-10-24, University of Texas at El Paso, 2010.

[9] A. G. Rivera and N. Ward, "Prosodic cues that lead to back-channel feedback in Northern Mexican Spanish," in *Proceedings of the Seventh Annual High Desert Linguistics Society Conference*, University of New Mexico, 2008.

[10] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, vol. 25, pp. 601–634, 2011.

[11] L. Huang, L.-P. Morency, and J. Gratch, "Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior," in *9th Int'l Conf. on Autonomous Agents and Multi-Agent Systems*, 2010.

[12] N. Ward and Y. Al Bayyari, "A case study in the identification of prosodic cues to turn-taking: Back-channeling in Arabic," in *Interspeech 2006 Proceedings*, 2006.

[13] T. K. Hollingsed and N. G. Ward, "A combined method for discovering short-term affect-based response rules for spoken tutorial dialog," in *Workshop on Speech and Language Technology in Education (SLaTE)*, 2007.

[14] K. P. Truong, R. Poppe, I. de Kok, and D. Heylen, "A multimodal analysis of vocal and visual backchannels in spontaneous dialogs," in *Interspeech*, pp. 2973–2976, 2011.

[15] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, pp. 1177–1207, 2000.

[16] A. Canavan and G. Zipperlen, *CALLHOME Mandarin Chinese Speech*. Linguistic Data Consortium, 1996. LDC Catalog No. LDC96S34, ISBN: 1-58563-080-2.

[17] D. Xudong, "The use of listener responses in Mandarin Chinese and Australian English conversations," *Pragmatics*, vol. 18, pp. 303–328, 2008.

[18] I. de Kok and D. Heylen, "A survey on evaluation metrics for backchannel prediction models," in *Interdisciplinary Workshop on Feedback Behaviors in Dialog*, 2012.

[19] D. Neiberg, "Visualizing prosodic densities and contours: Forming one from many," *TMH-QPSR (KTH)*, vol. 51, pp. 57–60, 2011.