UTEP Technical Report UTEP-CS-10-24

Visualization to Support 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
500 West University Avenue
El Paso, TX 79968-0518

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

August 25, 2010

Some meaningful prosodic patterns can be usefully represented with pitch contours, however the development of such descriptions is a labor-intensive process. To assist in the discovery of contours, visualization tools may be helpful. [Edlund et al., 2009] presented the idea of superimposing hundreds of pitch curves from a corpus as a way to see the overall patterns. In this paper we refine and extend this method and illustrate its utility in the discovery of a prosodic cue to back-channels in Chinese. We also discuss issues in relating a contour-based description to one in terms of a conjunction of featues, and illustrate this with a prosodic cue to back-channels in Spanish.

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

1 Why Contours?

In human dialog, turn-taking is largely managed by means of prosodic signals, or cues, exchanged by the participants. If a dialog system can correctly recognize and respond to these cues, it can make the user experience more efficient and more comfortable [Gratch et al., 2007, Raux and Eskenazi, 2009, Skantze and Schlangen, 2009].

A salient aspect of these cues is that they often seem to involve pitch contours, sometimes also called tunes: specific patterns of ups and downs over time. Such contours can can be drawn in various forms (Figure 1).

However, in the spoken dialog systems community, those working on turn-taking generally do not use contours, either explicitly or implicitly. Rather the "direct approach"

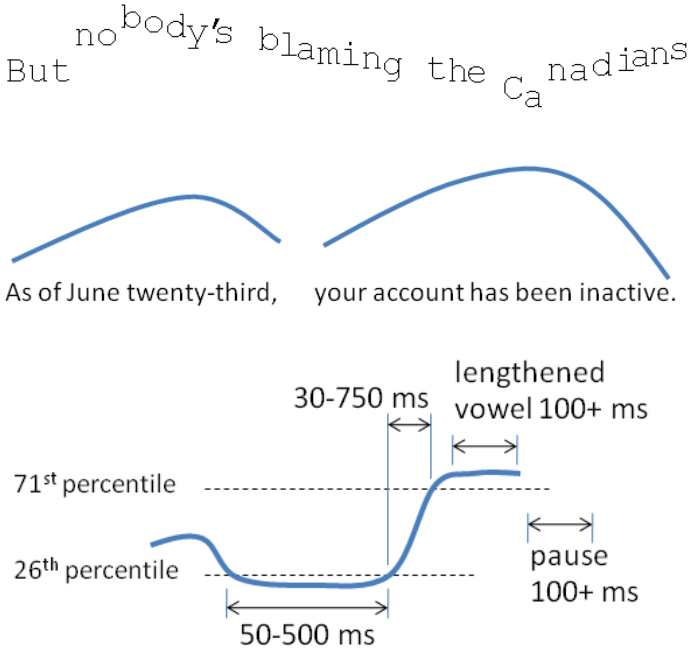

Figure 1: Examples of pitch contours: a contradiction contour (after [Bolinger, 1986] pg 246), a nonfinal contour followed by a final contour (after [Cohen et al., 2004], pg 183), and a back-channel cuing contour for Spanish (original work, building on [Rivera and Ward, 2008], see also Figure 6).

[Shriberg and Stolcke, 2004] 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 been quite successful, not only for turn-taking, but also more generally.

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", although referring to different mid-level features, may not actually differ in fact; but if drawn as contours the similarity or identity will be obvious. As another example, when describing a pattern in terms of features the temporal dependencies may not be immediately apparent, 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 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 [Gravano and Hirschberg, 2009, Huang et al., 2010] may have a natural implementation in terms of contours: the similarity be-

tween an input pitch curve and the cue contour may be an easy way to estimate cue strength. Contour-based descriptions can be used equally well for recognition and production. Finally, contours and the parameters describing them may serve as useful higher-level features for classifiers, and so contour discovery may also help those using the direct modeling method [Morency et al., 2010].

However contours currently have one great disadvantage: the difficulty of finding 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. Although there are helpful tools and methods [Ward and Al Bayyari, 2006, Hollingsed and Ward, 2007, Rosenberg, 2010], it is still not easy. Elicitation and instrumental techniques that work for monolog are hard to apply to prosodic patterns that relate to dialog-specific phenomena such as the prosody of attitude, information-state, speaker and interlocutor cognitive state, and turn-taking. Of course any single utterance has a clear pitch contour, but going from examples to a general rule is not straightforward.

This paper addresses this problem by presenting and illustrating a visualization method to support the discovery of meaningful pitch contours.

2 Visualization with Bitmap Clusters

Our visualization method is a refinement of Edlund, Heldner and Pelcé's "bitmap clusters" [Edlund et al., 2009]. Their innovation was to superimpose many individual pitch contours to reveal a general pattern:

> 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

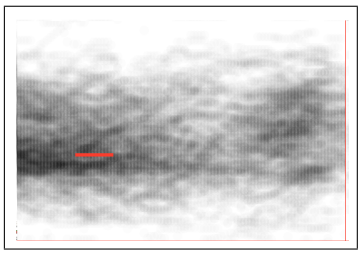

Figure 2: from Figure 6 of Edlund *et al.* [Edlund et al., 2009], by permission

They used this method to see what sort of contexts preceded various utterance types in the Swedish Map Task Corpus. Figure 2, taken from their paper, shows the contexts preceding 859 talkspurts which were tagged as 'very short utterances" and as having "low propositional content"; these 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.

3 Extensions and Refinements

We made a few improvements.

First, we chose the pitch regions to overlap in a different way. Edlund *et al.* aligned the ends of the talkspurts with the right edge of the display. This presumably reflects the assumption that the prosodic cues of interest occur at, and are aligned with, the 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. 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 [Ward and Tsukahara, 2000], 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, this does not require assumptions about the distribution, and thus may be more robust.

Third, we also 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 pass it off to some generic pre-processing phase that arbitrarily chops the input into utterances. Second, energy is important in identifying stressed syllables, fillers and so on.

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 or slow speaking rate. For pitch of course, the deltas are not always defined, so only when two adjacent pitch points are valid do we plot this. For both pitch and energy, what we plot is the difference between previous value, as a percentile, and the current pitch value, as a percentile.

Fifth, we extended the displays out to 2 seconds of past context, to look for longer-term patterns and to make it easier to see how the pitch contours in the cue region differ from the overall distributions.

Sixth, we did without pitch smoothing, not wanting to risk losing information. This is probably why the shading in our diagrams turned out more continuous and less blotchy than those of Edlund *et al.*

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 seen across the dialogs, we subtracted out the global average distribution, as estimated from a fairly random sample over the dialogs. (Specifically, the average value for each percentile in the leftmost 100ms of each display were calculated. Then these values were subtracted out from their respective percentiles over the entire display.) Before doing this the diagrams were blurry and hard to confidently interpret;

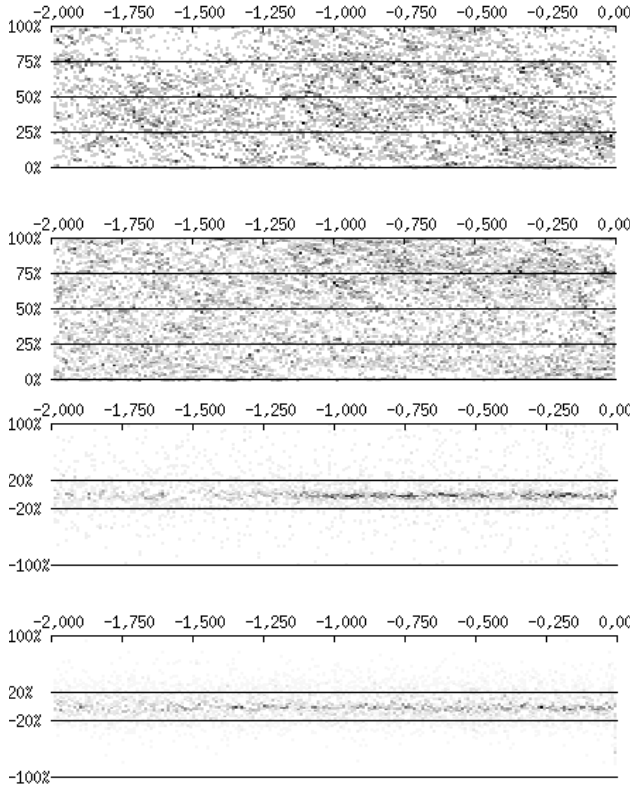

Figure 3: Overlaid Pitch, Energy, Delta Pitch and Delta Energy for Japanese

afterward they were much sharper, although somewhat more blotchy. Examples appear in the Appendix.

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

4 An Example

We developed these refinements in the course of trying visualizations for the contexts preceding back-channels in several languages. We chose to look at back-channeling because it is a well-studied 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 [Ward and Tsukahara, 2000].

While our hope was that these refinements to the visualization method would lead directly to clear diagrams with the prosodic cue immediately visible, sadly this was not the case. However it is possible to see useful information. For example: 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 from previous research: 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 found earlier looks somewhat different (requiring a region of pitch lower than the 26th percentile 460 to 350 ms before the back-channel onset [Ward and Tsukahara, 2000]) 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, as illustrated in the Appendix.

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 [Canavan and Zipperlen, 1996], 90 minutes in total, we had two native speakers independently identify all back-channels according to the criteria of [Ward and Tsukahara, 2000]. 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 labeler were *uh*, *oh*, *dui*, *uh-huh*, *em*, *shima*, *hmmm*, *ok*, *yeah*, *huh*, *duia*, *uhuh*, *shia*, *hmmmm*, and *good*.

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 [Ward and Tsukahara, 2000], 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 infrastrucure previously developed for extracting prosodic features and making similar decisions for other languages, but with no knowledge of Chinese, got to work.

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 at the level of 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 the 15th percentile and lasting at least 220ms, followed by

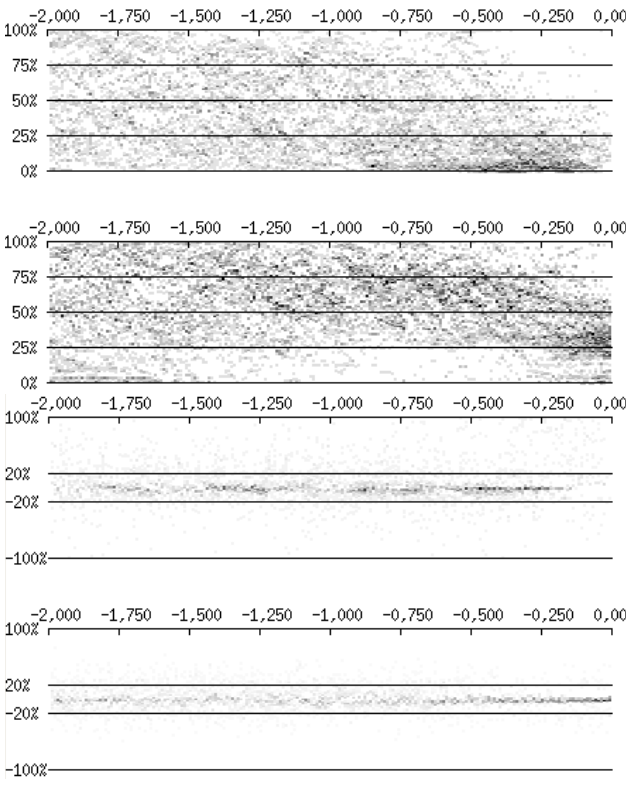

Figure 4: Overlaid Pitch, Energy, Delta Pitch and Delta Energy for Chinese

- 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 4% accuracy.

6 From Contours to Rules

Although contours may be fully adequate, as discussed in section 1, it is certainly also convenient, at least, to describe prosodic rules using a more symbolic representation, such as a conjoined set of clauses specifying what features must be present and how they must relate to each other in time. Among other things, such a representation can explicitly represent the degree to which the component features may stretch internally or relative to each other.

Going from a contour-based description to a formal rule is, however, not easy. For example, the contour shown in figure 1 for spanish has had two conversions to qualitative rules, that reported in [Rivera and Ward, 2008] and the new rule seen in figure 6. The new rule was developed in part because the code implementing the original rule was lost, but also because we wanted a rule whose implementation would be cleaner. An example of an unclean implementation would be one that first looks for one predictive feature, then hunts back and forth in the buffers and arrays looking for the presence of the additional features. A somewhat more elegant approach is to store all the necessary information at all times, so that, for example, at every frame there is an array

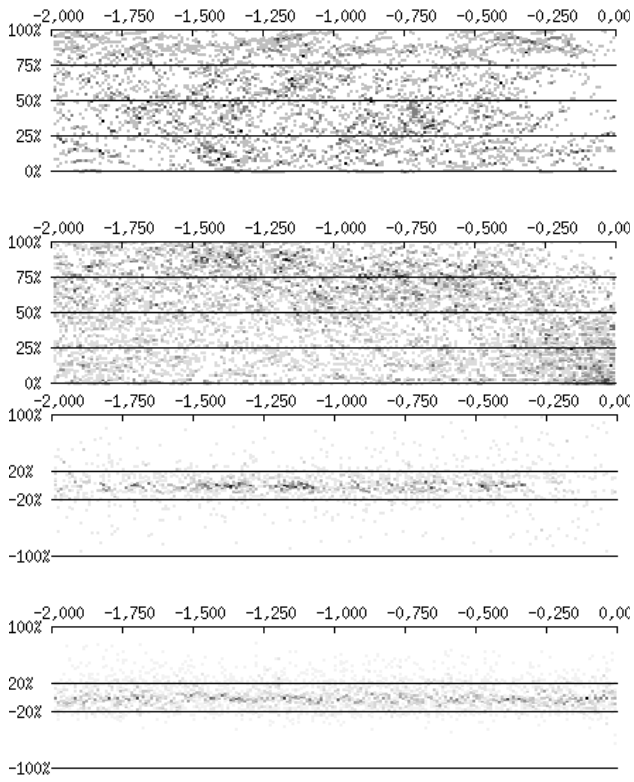

Figure 5: Overlaid Pitch, Energy, Delta Pitch and Delta Energy for Spanish

value for every possible low-level feature of interest (current-frame-is-low-pitch, currently-in-a-low-pitch-region, time-stince-start-of-low-pitch-region, etc.).

Possibly the most elegant (or at least, efficient in terms of computation and memory) is to implement a rule as a finite state machine that operated on-line, processing the speech input frame by frame. This new rule was relatively straightforward to implement in this way, as seen in Figure 7. This new rule gives a coverage of 26% and an accuracy of 13%.

7 Future Work

From the experience discovering the Chinese pre-backchannel pattern, we conclude that this visualization method has value.

However it clearly has room for improvement. Consider Figure 5, displaying the contexts of 152 back-channels in Spanish [Rivera and Ward, 2008]. Some things are evident, including a dearth of high-pitch points, a clear quiet region pause in the last quarter second, and possibly a tendency for flat pitches from -1500 to -1000, as seen from the deltas. However there is not much else to see, even knowing the pattern we expect to find.

There are several possible ways to improve these visualizations. One could explicitly display duration or rate. One could make the features more robust, for example by computing the energy deltas over frames wider than 10 ms. One might apply a thinning algorithm to visually accentuate the tendencies, to turn cloudy streaks into nice curves. Finally, one could improve the way the

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

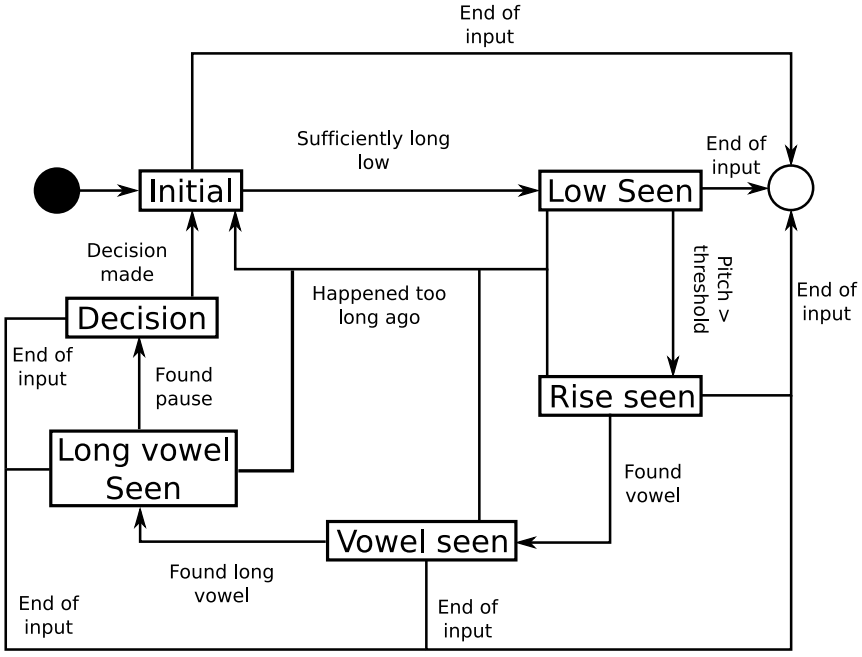

Figure 7: Finite state machine for making back-channel predictions for Spanish

horizontal alignment is done in generating the overlays. As reaction times vary, the time from the prosodic cue, whatever it may be, to the response will not be constant. It may be possible to devise an expectation-maximization algorithm, where the horizontal alignments are iteratively adjusted to make the pitch contours align better.

Although contour-based descriptions have limitations, their discovery will be easier with this new tool, the overlaid prosodic display, and this may also be a generally useful addition to the prosodic analysts' toolbox.

References

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

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

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

[Edlund et al., 2009] Edlund, J., Heldner, M., and Pelcé, A. (2009). Prosodic features of very short utterances in dialogue. In *Nordic Prosody - Proceedings of the Xth Conference*, pages 56–68.

[Gratch et al., 2007] Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., van der Werf, R., and Morency, L. (2007). Can Virtual Humans Be More Engaging Than Real Ones? *Lecture Notes in Computer Science*, 4552:286–297.

[Gravano and Hirschberg, 2009] Gravano, A. and Hirschberg, J. (2009). Backchannel-inviting cues in task-oriented dialogue. In *Interspeech*, pages 1019–1022.

[Hollingsed and Ward, 2007] Hollingsed, T. K. and Ward, N. G. (2007). 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)*.

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

[Morency et al., 2010] Morency, L.-P., de Kok, I., and Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20:70–84.

[Raux and Eskenazi, 2009] Raux, A. and Eskenazi, M. (2009). A finite-state turn-taking model for spoken dialog systems. In *NAACL HLT*.

[Rivera and Ward, 2008] Rivera, A. G. and Ward, N. (2008). 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.

[Rosenberg, 2010] Rosenberg, A. (2010). Classification of prosodic events using quantized contour modeling. In *HLT-NAACL 2010*, pages 721–724.

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

[Skantze and Schlangen, 2009] Skantze, G. and Schlangen, D. (2009). Incremental dialogue processing in a micro-domain. In *EACL*, pages 745–753.

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

[Ward and Al Bayyari, 2007] Ward, N. and Al Bayyari, Y. (2007). A prosodic feature that invites back-channels in Egyptian Arabic. In Mughazy, M., editor, *Perspectives on Arabic Linguistics XX*, pages 186–206. John Benjamins.

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

[Ward et al., 2006] Ward, N. G., Novick, D. G., and Salamah, S. I. (2006). The utep corpus of iraqi arabic. Technical Report UTEP-CS-06-02, University of Texas at El Paso, Department of Computer Science.

Appendix

Section 3 mentioned that the average distributions were subtracted out, in order to "clear up" the display and reveal more patterns. This appendix shows diagrams for three additional languages, first to show the value of subtracting the means, and second to provide futher illustrations of the information that can be seen in such diagrams.

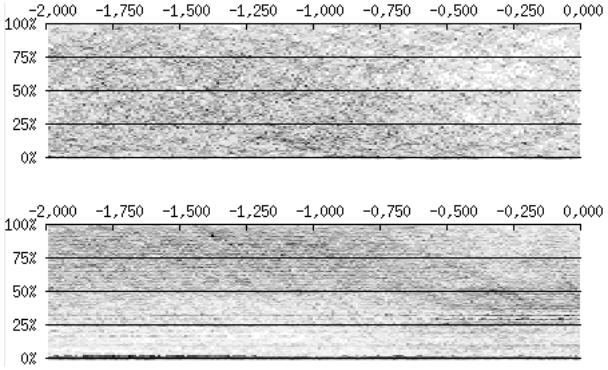

Figure 8: Pitch (top) and Energy (bottom) for Egyptian, without subtracting means

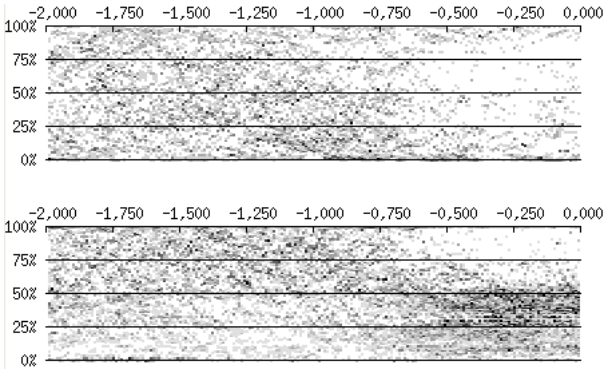

Figure 9: Pitch and Energy for Egyptian, after subtraction

For Egyptian Arabic, the various visualizations (Figures 8–11, based on the contexts of 393 back-channels) [Ward and Al Bayyari, 2007] are informative. The pitch tends to drop off at around –600ms and the energy also drops down below the 60th percentile at –600ms. There is an indication of a flat pitch region from –2500ms to –750ms in the delta pitch plot, and delta energy is fairly flat from –800ms up until the backchannel.

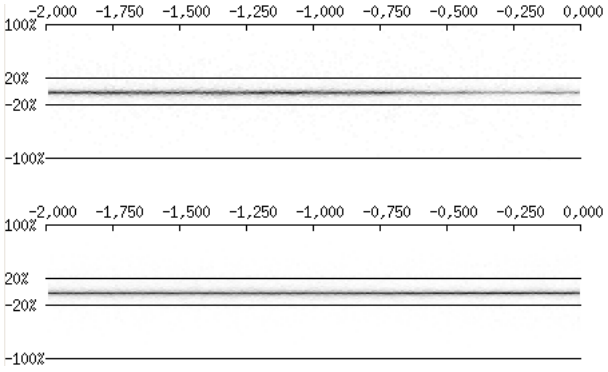

Figure 10: Delta Pitch and Energy for Egyptian, without subtraction

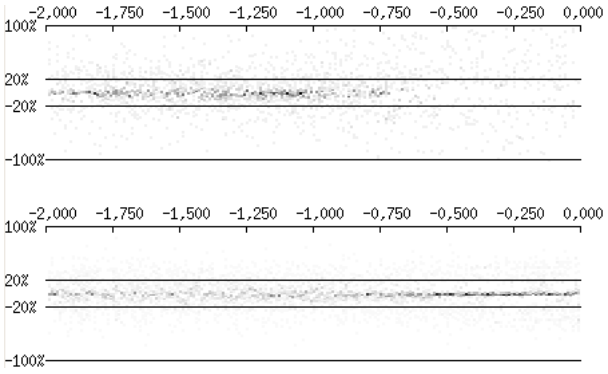

Figure 11: Delta Pitch and Energy for Egyptian, after subtraction

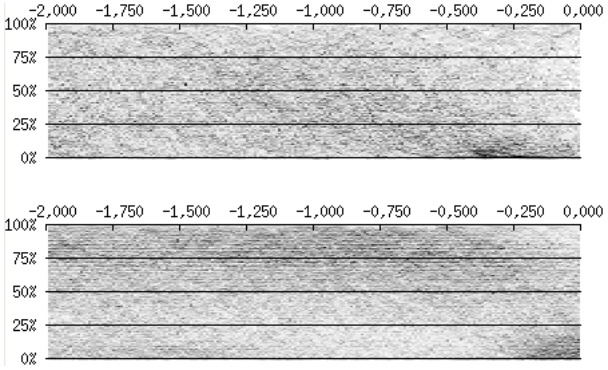

Figure 12: Iraqi Arabic, without means subtracted, Pitch and Energy

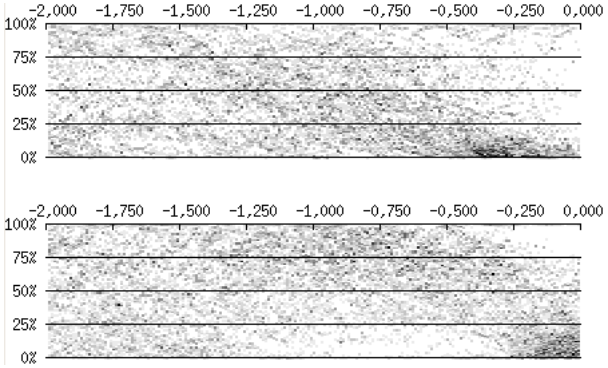

Figure 13: Iraqi Arabic, with means subtracted

For Iraqi Arabic, in Figures 12–15, based on 556 instances [Ward et al., 2006, Ward and Al Bayyari, 2006], we see an indication of a low pitch region below the 20th percentile from –500ms to –250ms, with fewer pitch points thereafter. The energy is fairly high around –1000ms to –500ms, but it then drops below the 25th percentile until the backchannel, indicating that there is typically a pause before the other person backchannels; a tendency that, incidentally, was more extreme in the Egyptian Arabic data, although that may be due to the fact that that was telephone data.

The delta pitch shows a consistent region of pitch from –1,300ms to –400ms and a very flat region from –400ms to –250ms. After that, the pitch drops off. Delta energy seems to cluster mostly between 10 and –10 throughout until the backchannel.

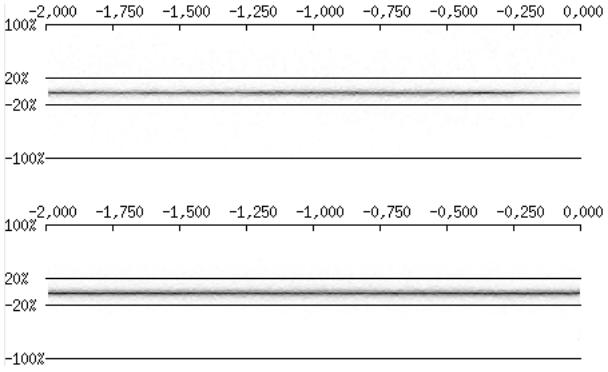

Figure 14: Iraqi Arabic Deltas, without means subtracted, Pitch and Energy

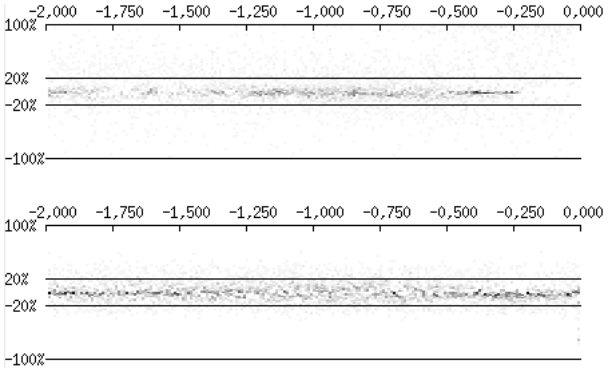

Figure 15: Iraqi Arabic Delta, with means subtracted

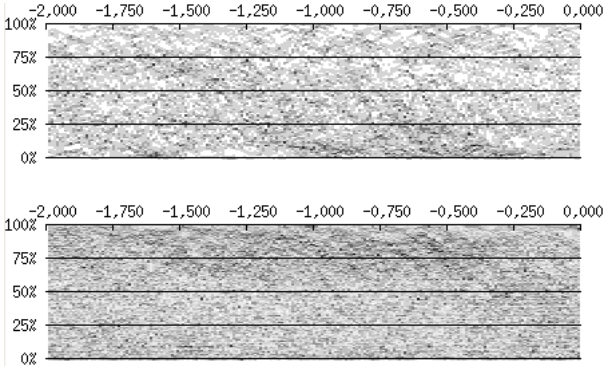

Figure 16: Pitch and Energy for American English, without subtracting means

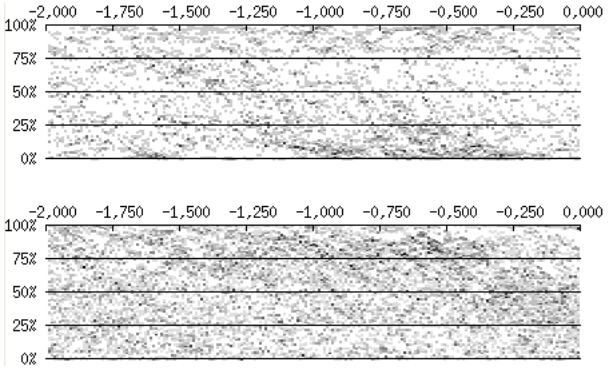

Figure 17: Pitch and Energy for American English, after subtraction

For English, based on 309 back-channels, we see the expected low pitch region [Ward and Tsukahara, 2000], here below the 40th percentile, from around –1000ms to –500ms. For energy, it appears there is a high region from –1600ms to –300ms which then drops below the 60th percentile around –300ms. Delta pitch seems to indicate a pitch drop around –500ms.

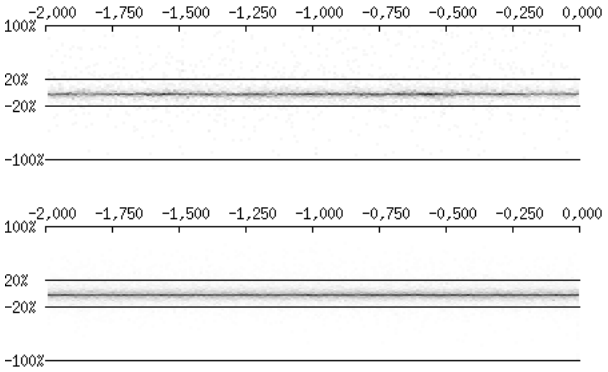

Figure 18: Delta Pitch and Energy for American English, without subtraction

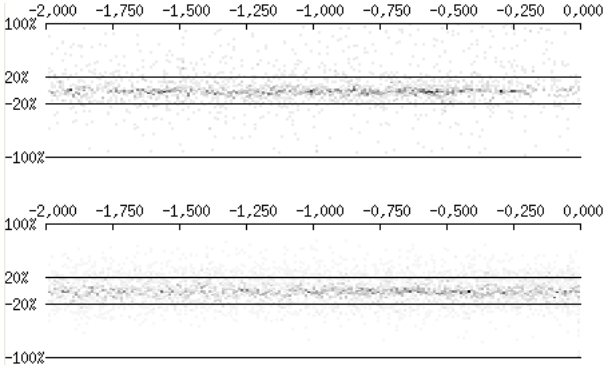

Figure 19: Delta Pitch and Energy for American English, after subtraction

Thus it seems clear that subtracting the means helps to reduce the amount of noise in each display, revealing patterns that previous work had indicated we would find, and occasionally new ones. The delta displays generally were uninformatively constantly flat until the subtractions were done, but after subtracting the means, suggestive patterns sometimes became visible there too.

Acknowledgment

This work was supported in part by the NSF as Project No. 0415150 and by RDECOM via USC ICT. We thank an anonymous Interspeech referee for comments.