Entrainment analysis of categorical intonation representations

Uwe D. Reichel[1], *Jennifer Cole*[2]

[1]Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
[2]Department of Linguistics, University of Illinois, United States of America
uwe.reichel@nytud.mta.hu, jscole@illinois.edu

Abstract

Most studies on prosodic entrainment focus on coarse parametric variables as f0 mean and standard deviation. Only recently first attempts were made to measure entrainment also for categorical intonation representations namely pitch accent types [1]. We propose further metrics for this purpose adopted from text similarity measurement and alignment. These metrics were applied to quantify the similarity of automatically derived intonation contour class sequences in cooperative and competitive dialogs. In line with previously reported results for parametric variables we found also for the categorical representation higher similarities and thus more entrainment in the cooperative dialogs than in the competitive ones. The introduced metrics can be of use for any entrainment research on categorical data as e.g. for ToBI label sequences.

Index Terms: entrainment, intonation stylization, string similarity, local alignment

1. Introduction

In conversation speakers accommodate more and more to each other. This phenomenon is called entrainment and can be observed on various phonetic and linguistic levels. On the linguistic level entrainment affects amongst others the choice of words [2] and syntactic constructions [3, 4]. On the phonetic level entrainment was revealed in dialog data and shadowing experiments for speaking rate [5, 6], intensity [5, 6], voice quality [6], and pitch [7, 8, 5, 9]. Entrainment turned out to be stronger in case of mutual positive attitude of the interlocutors, than in case of negative attitude [10]. Furthermore, entrainment has been shown to increase the success of conversation in terms of low inter-turn latencies and a reduced number of interruptions [6, 2]. Consequently, more entrainment has been reported in cooperative than in competitive dialogs e.g. with respect to intonation contour shapes [11]. These findings are in line with theoretical models such as the Communication Accommodation Theory [12] stating that entrainment enhances social approval and communication efficiency.

For intonation entrainment research is so far mostly restricted to parametric variables, most of them coarse as for example f0 mean and standard deviation over utterance stretches. Only few attempts have so far been made to measure entrainment for a higher-level categorical intonation representation. [1] measured global entrainment over entire dialogs in terms of perplexity and Kullback-Leibler divergence on ToBI [13] pitch accent and boundary tone trigrams. Furthermore, they addressed local entrainment in temporally closely related speech chunks using the Levenshtein distance between tone sequences.

This study aims to contribute to these new entrainment analyses of categorical intonation representations in the following way:

- It will be shown, how such a representation can be generated in a bottom-up way (section 3).
- We will introduce similarity measures for this representation, that capture local entrainment within neighboring speech chunk pairs.
- These measures provide a better account to sub-sequence and crossing alignments of tone sequences than does a Levenshtein distance based approach (section 4).

The employed similarity metrics are: Jaccard index, Cosine index, Szymkiewicz-Simpson coefficient, as well as a similarity measure derived from local alignment.

We applied these metrics to cooperative and competitive dialog data (section 2) to see whether the found similarity values are in line with the findings on parametric data mentioned above. Concretely, we hypothesize to find more entrainment in cooperative than in competitive dialogs expressed by higher values of all proposed similarity metrics.

2. Data

We used parts of the Illinois Game Corpus [14] that contains *Tangram* game dialogs by American English speakers in cooperative and competitive settings. The tangram is a puzzle consisting of seven pieces that can be combined to various shapes. Both dialog partners were separately presented with Tangram silhouettes that were reciprocally hidden from the view of the other partner. The task was to decide whether the silhouettes are the same or different by verbally describing them to each other. In the cooperative setting the partners solved this common goal in a joint effort. In the competitive setting, the partners were required to solve this task competitively, and the one solving it first was declared to be the winner. For more details about the recording setting please consult [15]. For the current study a subset of ten dialogs by five interlocutor pairs was used, of which three were Female-Female pairs and two were Male-Female pairs. Each interlocutor pair took part in a cooperative and a competitive condition, thus our data comprises paired samples of five cooperative and competitive dialogs. Mean dialog duration amounts to 6.5 minutes.

The dialogs were manually text-transcribed and chunk-segmented, and partly manually dialog-act annotated using the tag set of [16]. The data was signal-text aligned by the WEBMAUS webservice [17, 18] and was part of speech tagged using the Balloon toolkit [19]. Both alignment and part of speech labels serve to automatically locate prosodic events, i.e. phrase boundaries and potential pitch accent locations as described in [20].

F0 was extracted by autocorrelation (PRAAT 5.3.16 [21], sample rate 100 Hz). Voiceless utterance parts and F0 outliers were bridged by linear interpolation. The contour was then

smoothed by Savitzky-Golay filtering using third order polynomials in 5 sample windows and transformed to semitones relative to a base value [22]. This base value was set to the F0 median below the 5th percentile of an utterance and serves to normalize F0 with respect to its overall level.

3. Categorical intonation representation

For intonation stylization we adopt the parametric CoPaSul approach of [20], which is illustrated in Figure 1. Within this framework intonation is stylized as a superposition of linear global contours, and third order polynomial local contours. The domain of global contours approximately related to intonation phrases is determined automatically by placing prosodic boundaries at speech pauses and punctuation in the aligned transcript. The domain of local contours is determined by placing boundaries behind each content word so that the resulting segments generally contain at most one pitch accent.

The global linear component is given by the F0 baseline fitted through f0 values at the bottom of the time varying f0 range as explained in [23]. The baseline is then subtracted from the F0 contour, and a third order polynomial is fitted to the F0 residual within each local segment. Time is normalized to the range from -1 to 1 so that time 0 is placed in the mid of the content word's syllable bearing the lexical stress.

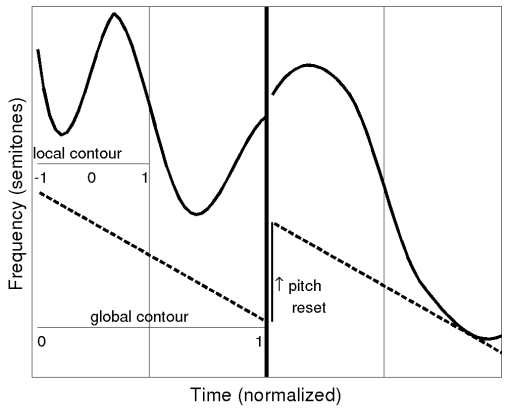

Figure 1: *CoPaSul: Contour-based parametrical superpositional F0 stylization.*

In order to derive a categorical representation from this parametric stylization, the slopes of the global contours as well as the polynomial coefficients of the local contours are clustered by Kmeans. Following [20] the optimal number of contour classes was initialized by subtractive clustering [24]. The resulting three global and four local contour classes are shown in Figure 2.

4. Entrainment measurements

As one can see in Figure 3, the contour class distributions, unigrams as well as bigrams, are highly determined by the dialog act of the speech chunk. This is reflected by significantly higher information radii (two-sided Welch tests, $p < 0.001$) of these distributions when comparing them between different dialog act chunks as opposed to same dialog act chunks. These findings are in line with [25] who discuss dialog-related differences in intonation parameters in the context of Ohala's Frequency Code

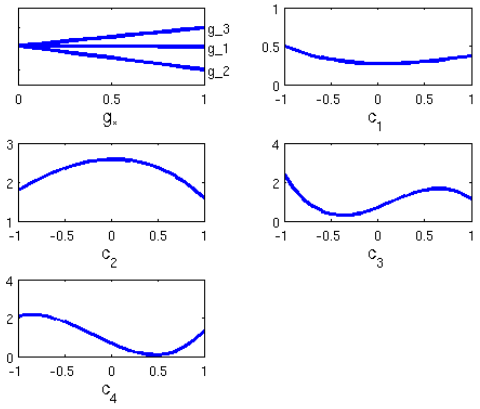

Figure 2: *Global (g_i) and local (c_j) contour classes resulting from polynomial coefficient clustering.*

framework [26]. In order to disentangle entrainment and dialog act dependencies, we applied the similarity measures only on speech chunks of the same dialog act.

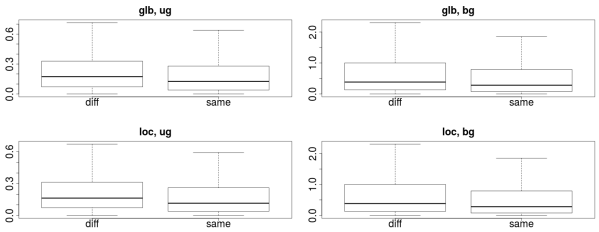

Figure 3: *Information radii of contour class unigram and bigram probability models within and across dialog act types.*

4.1. Similarity of contour class inventories

The similarity of the contour class inventories X and Y of speech chunk pair was quantified by three standard string-based similarity metrics [27]: the Cosine similarity, the Jaccard index [28] and the overlap ratio (Szymkiewicz-Simpson coefficient [29]), which are defined as follows:

$$\text{Cosine } C(X,Y) = \frac{|X \cap Y|}{\sqrt{|X||Y|}},$$

$$\text{Jaccard } J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|},$$

$$\text{Overlap } O(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)}.$$

$|S|$ refers to the cardinality of a set S, i.e. in our case the number of different contour class types. All indices range from 0 (no similarity) to 1 (total similarity).

4.2. Similarity of contour class sequences

We adopted the idea of [1] to measure similarity of contour class sequences by means of alignment. Since sequences usually differ in length, and since these length differences add up to the

overall distance, it is advisable to normalize the distance with respect to length. [1] propose the following transformation of the Levenshtein distance $d(x, y)$ between the sequences x and y to a similarity score $s_r(x, y)$ ranging between 0 and 1 partly normalized with respect to length:

$$s_r(x, y) = \frac{m - d(x, y)}{m},$$

where $m = \max[\text{length}(x), \text{length}(y)]$, i.e. the length of the longer sequence and thus the upper limit of the number of edit operations. Note that x and y here do not refer to sets as the capital letters in the previous section, but to contour class sequences. As one can see in Figure 4, this similarity measure has two shortcomings: first, it does punish sequences of different length even if one sequence is entirely contained within the other. Thus two possible domains of entrainment, utterance duration and intonation, are merged to a single metrics. Second, it does punish sequences with cross matching subsequences. Thus, it cannot account for cases where interlocutors choose the same intonation contours but at different positions within their utterances. To disentangle duration and intonation and to capture cross matches we propose an alternative measure based on local alignment:

$$s_l(x, y) = \frac{\text{length(localigned(z))}}{\text{length}(z)},$$

where $z = \arg\min_{z \in \{x,y\}}[\text{length}(z)]$. The similarity $s_l(x, y)$ of an intonation class sequence pair is thus the proportion of the locally aligned parts of the shorter sequence in that pair. As s_r also s_l similarity scores range from 0 to 1, Figure 4 gives an example. Since all members of the shorter sequence x are (with cross matches) contained in the longer sequence y, $s_l(x, y)$ amounts 1. In contrast, the Levenshtein distance between x and y amounts 6 which yields a similarity $s_r(x, y) = \frac{7-6}{7} = 0.14$, and thus a quite different result, that underestimates the fact, that x is entirely contained in y.

x	**e**	**f**	**g**	**a**	**b**		
y	**a**	**b**	c	d	**e**	**f**	**g**

Figure 4: Alignment of two sequences x and y of differing length. x is with cross correspondences entirely contained in y. Levenshtein distance: 6; Levenshtein-derived similarity $s_r(x, y) = 0.14$; local alignment derived similarity $s_l(x, y) = 1$.

The proposed local alignment is implemented by an adaptation of the dynamic programming Smith-Waterman algorithm [30]. The alignment score matrix H spanned by the sequences x and y with length m and n, respectively (cf. left half of Figure 5) is filled as follows:

$$H[i, 0] = 0, 0 \le i \le m$$
$$H[0, j] = 0, 0 \le j \le n$$
$$H[i, j] = \max \left\{ \begin{array}{rcl} 0 & : & \text{Lower bound} \\ H[i-1, j-1] + s(x_i, y_j) & : & \text{Match/Mismatch} \\ \max_{k>0}[H(i-k, j) + W_k] & : & \text{Deletion} \\ \max_{l>0}[H(i, j-l) + W_l] & : & \text{Insertion} \end{array} \right\},$$
$$1 \le i \le m, 1 \le j \le n$$

$s(a, b)$ is a similarity function and W_i a gap scoring scheme [31]. Both allow for a high flexibility in the alignment process. For our purpose we restrict it to align only matching subsequences. Thus everything but zero-substitutions should result in a cell value below or equal 0 so that this operation will not contribute to the alignment. This is realized by setting W_i as well as $s(a, b)$ for $a \neq b$ constant to $-l$, where l is the length of any of the sequences to be aligned. Only zero-substitutions ($a == b$) are rewarded by $s(a, b) = 1$.

All matching subsequences are then retrieved from this matrix by the following iteration:

while $\max(H) > t$

– trace back from the cell containing this maximum the path leading to it until a zero-cell is reached
– add the subsequence collected on this way to the set of aligned sequences
– set all traversed cells to 0

This iteration is illustrated in Figure 5. The threshold t defines the required minimum length of aligned subsequences. It is set to 2 in this study. $t = 1$ would result in a complete alignment of any pair of permutations of x. The traversed cells need to be set to 0 after each iteration step to prevent that one subsequence would be related to more than one alignment pair.

This approach allows for two more restrictions: to prevent cross alignment not just the traversed cells $[i, j]$ but for each of these cells its entire row i and column j needs to be set to 0. Second, if only the longest common substring is of interest, then the iteration is trivially to be stopped after the first step.

	-	a	b	c	d	**e**	**f**	**g**
-	0	0	0	0	0	0	0	0
e	0	0	0	0	0	**1**	0	0
f	0	0	0	0	0	0	**2**	0
g	0	0	0	0	0	0	0	**3**
a	0	1	0	0	0	0	0	0
b	0	0	2	0	0	0	0	0

	-	**a**	**b**	c	d	**e**	**f**	**g**
-	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0
a	0	**1**	0	0	0	0	0	0
b	0	0	**2**	0	0	0	0	0

Figure 5: Iterative longest common subsequence (LCS) detection in local alignment. While the matrix maximum is above a threshold, start at this maximum and trace back until a 0 cell is reached and set all traversed cells to 0. This yields in the first iteration step (**left**) the alignment of **e f g**, and in the second step (**right**) the alignment of **a b**.

5. Results

In line with mentioned findings of previous studies and with our hypothesis all similarity measures yield higher values in the cooperative than in the competitive dialogs (two-sided Welch tests, $p < 0.001$). This is shown in Figure 6.

6. Discussion

We introduced several similarity metrics from natural language processing to measure entrainment in categorical intonation data. The results indicate higher entrainment for both intonation inventory as well as tone sequencing which is well in line with finding on the parametric level. This we take as an indication that the proposed metrics are of value in prosodic entrainment research. We argue that local alignment based similarity is better suited for entrainment measurements than the transformed standard Levenshtein distance since it cancels out sequence length differences and can cope with cross correspondences. It is highly flexible due to several tuning parameters

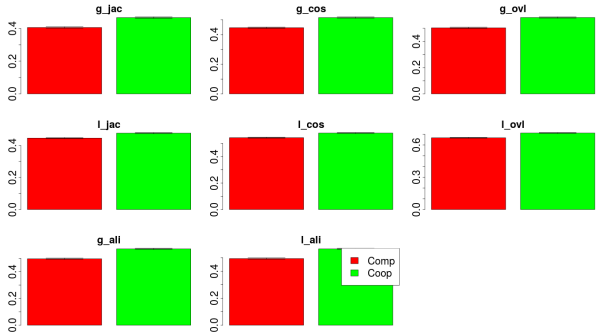

Figure 6: *Similarities of global (g_*) and local (l_*) contour class inventories in competitive (COMP, red) and cooperative (COOP, green) dialogs. jac – Jaccard index, cos – cosine similarity, ovl – overlap ratio, ali – local alignment.*

given by the similarity function, the gap penalty scoring, the score thresholding, and the procedure how to trace back the alignment score matrix, so that it can be customized to the respective research needs.

In this study the categorical intonation representation was derived in a bottom-up way. Nevertheless, the measures can be applied to any categorical data including expert-driven intonation representations as ToBI annotations.

7. Acknowledgments

The work of the first author is financed by a grant of the Alexander von Humboldt society.

8. References

[1] A. Gravano, v. Beňuš, R. Levitan, and J. Hirschberg, "Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement," in *Proc. Interspeech*, Dresden, 2015, pp. 578–582.

[2] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, 2008, pp. 169–172.

[3] A. Cleland and M. Pickering, "The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure," *Journal of Memory and Language*, vol. 49, pp. 214–230, 2003.

[4] S. Gries, "Syntactic priming: A corpus-based approach," *Journal of Psycholinguistic Research*, 2005.

[5] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 3081–3084.

[6] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *NAACL HLT '12 Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, 2012, pp. 11–19.

[7] S. Gregory and S. Webster, "A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions," *J. Pers. Soc. Psychol.*, vol. 70, pp. 1231–1240, 1996.

[8] S. Gregory, K. Dagan, and S. Webster, "Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality," *J. Nonverbal Behavior*, vol. 21, pp. 23–43, 1997.

[9] M. Babel and D. Bulatov, "The role of fundamental frequency in phonetic accommodation," *Language and Speech*, vol. 55, pp. 231–248, 2011.

[10] C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proc. Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 793–796.

[11] J. Cole and U. Reichel, "Prosodic entrainment – the cognitive encoding of prosody and its relation to discourse function," Keynote at Framing speech satellite workshop of the Speech Prosody conference, Boston, 2016.

[12] H. Giles and N. Coupland, *Language: Contexts and Consequences*. Pacific Grove, CA: Brooks/Cole, 1991.

[13] J. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, MIT, Cambridge, Massachusets, 1980.

[14] PAGE, "Prosodic and Gestural Entrainment in Conversational Interaction across Diverse Languages," http://page.home.amu.edu.pl/.

[15] U. Reichel, N. Pörner, D. Nowack, and J. Cole, "Analysis and classification of cooperative and competitive dialogs," in *Proc. Interspeech*, Dresden, Germany, 2015, p. paper 3056.

[16] J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson, "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, vol. 23, no. 1, pp. 13–31, 1997.

[17] F. Schiel, "Automatic Phonetic Transcription of Non-Prompted Speech," in *Proc. ICPhS*, San Francisco, 1999, pp. 607–610.

[18] T. Kisler, U. Reichel, F. Schiel, C. Draxler, B. Jack l, and N. Pörner, "BAS Speech Science Web Services - an update of current deve lopments," in *Proc. LREC 2016*, Portoro, Slovenia, 2016, pp. 3880–3885.

[19] U. Reichel, "PermA and Balloon: Tools for string alignment and text processing," in *Proc. Interspeech*, Portland, Oregon, USA, 2012, p. paper no. 346.

[20] ——, "Linking bottom-up intonation stylization to discourse structure," *Computer, Speech, and Language*, vol. 28, pp. 1340–1365, 2014, dOI: 10.1016/j.csl.2014.03.005.

[21] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer," Institute of Phonetic Sciences of the University of Amsterdam, Tech. Rep., 1999, 132–182.

[22] A. Savitzky and M. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[23] U. Reichel and K. Mády, "Comparing parameterizations of pitch register and its disconti nuities at prosodic boundaries for Hungarian," in *Proc. Interspeech 2014*, Singapore, 2014, pp. 111–115.

[24] S. Chiu, "Fuzzy Model Identification Based on Cluster Estimation," *Journal of Intelligence & Fuzzy Systems*, vol. 2, no. 3, pp. 267–278, 1994.

[25] K. Mittelhammer and U. Reichel, "Characterization and prediction of dialogue acts using prosodic features," in *Elektronische Sprachverarbeitung 2016*, ser. Studientexte zur Sprachkommunikation, O. Jokisch, Ed. Dresden, Germany: TUDpress, 2016, vol. 81, pp. 160–167.

[26] J. Ohala, "The frequency code underlies the sound symbolic use of voice pitch," in *Sound Symbolism*. Cambridge: Cambridge University Press, 1994.

[27] W. Gomaa and A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, 2013.

[28] P. Jaccard, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines," *Bulletin de la Socit Vaudoise des Sciences Naturelles*, vol. 37, pp. 241–272, 1901.

[29] D. Szymkjewicz, "Une contribution statistique a la géographie floristique," *Acta Soc. Bot. Polon.*, vol. 34, no. 3, pp. 249–265, 1934.

[30] T. Smith and M. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.

[31] M. Vingron and M. Waterman, "Sequence alignment and penalty choice. Review of concepts, case studies and implications," *Journal of Molecular Biology*, vol. 235, no. 1, pp. 1–12, 1994.