MAPPING INTONATIONAL TRANSCRIBERS' TONE SIMILARITY SPACE

Rebecca Herman and Julia Tevis McGory
Department of Linguistics, The Ohio State University

ABSTRACT

This paper discusses a newly-developed method for comparing the factors used by different prosodic labelers in intonational transcription. Four experienced intonational transcribers' tone similarity judgments were elicited, and these judgments were submitted to a multi-dimensional scaling analysis. From this analysis, we can determine dimensions and tonal clustering in each transcriber, and the subjective similarity space for the categories in an intonational transcription system. This information is crucial for studies which measure inter-transcriber reliability, in which it is important to determine not just whether transcribers differ from each other but also how they differ.

1. INTRODUCTION

Prosodic transcription is being used currently to produce databases of prosodically labeled speech, which are useful for investigations of intonational phenomena. Prosodic transcription involves labeling of relatively high and low tones. In the process of identifying tonal events in the signal, many factors are taken into consideration, including the shape and level of the F0 contour in the region under consideration, the perceived stress of the syllable associated with these F0 changes, and the degree of disjuncture between words.

ToBI (Tones and Break Indices) is a labeling standard used for the transcription of English intonation, including mainstream varieties of American English [1]. We will be concentrating on the tonal level, which describes movements between relatively low (L) and high (H) fundamental frequency (F0). These L and H elements (as well as combinations of them) make up the inventory of pitch accents (associated with stressed syllables) and edge tones (aligned with the right edge of a phrase). There are two mono-tonal pitch accents, represented by H* and L*. There are two rising bi-tonal pitch accents, represented by L+H* and L*+H, where the * indicates association of the starred tone with the stressed syllable and the + indicates that the other tone is leading or trailing, but not associated to a particular syllable. Downstep also occurs in English intonation, represented in the set of pitch accents by !H*, L+!H*, L*+!H, as well as H+!H*. This gives a total of 8 pitch accents in the inventory. L and H elements also make up the inventory of phrase accents and boundary tones. Phrase tones, which delineate intermediate phrases, are represented by H- or L-. There is also the possibility of a downstepped phrase accent in English intonation, represented by !H-. Boundary tones, which delineate intonational phrases, are represented by H% or L%.

Because of the many factors involved in labeling and because of the many possibilities in types of tones in the ToBI transcription system, we were curious to examine which dimensions affect transcribers' choices of tone labels and the relative importance of each dimension.

For example, an F0 movement from a low to a high value somewhere in an utterance could be analyzed in a number of ways. Some possibilities include analyzing this as a single, bitonal pitch accent, with both L and H components. These would be L+H* or L*+H, differing in which tone is anchored to the stressed syllable. Alternatively, the low portion could be analyzed as a L* pitch accent anchored to a syllable with phrase level stress, with the high portion being a H- phrase accent (resulting in L*H-). Yet another alternative is to analyze the low portion by labeling it as a L- phrase accent, with the high portion being a H% boundary tone (L-H%). What would differentiate between these analyses would be, for example, the perception of prominence in the syllable with the low F0, which would indicate that the low portion of the F0 could be accounted for by a pitch accent containing a L* (L*+H or L*). The absence of such a percept of prominence in the low F0 area would lead the transcriber to account for the low F0 by a leading L+ tone in a bitonal L+H* pitch accent. If there were a subjective sense of disjuncture after this low F0 area which implied a phrase break, then this region of low F0 could be accounted for by a L- phrase accent followed by a H%. Thus, different analyses of a particular tonal event can be chosen on the basis of a number of dimensions.

The factors described above not only govern transcribers' decisions during labeling, but also come into play in defining which tones are similar for a transcriber. For example, a L+H* could be considered to be similar to a H* along several dimensions: both accents are quite prominent, involve a high pitch level, and may have very similar pitch shapes in phrase-initial position. Thus, one question is whether transcribers have similar subjective impressions of tone similarities. Using the type of generalization stated above, Pitrelli et al. [2] merged certain tonal categories in their inter-transcriber reliability analysis. Nakatani [3] and Jannedy [4] also merged tonal categories in their analyses.

This paper aims to evaluate such intuitions about similarity in tones by eliciting tone similarity judgments from several experienced intonational transcribers. We use these similarity judgments in a multi-dimensional scaling analysis (MDS) in order to observe the dimensions used by different transcribers as well as to find clustering patterns of tones within different transcribers' tone similarity spaces. (This study expands on work by Herman [5].)

2. METHODS

2.1. Subjects

We elicited tone similarity judgments from four transcribers, experienced in ToBI labeling. All of the subjects had at least a year of experience in ToBI transcription and are native English speakers.

2.2. Elicitation of Tone Similarity

The 4 subjects completed a questionnaire consisting of a list of comparisons between pairs of tones. This was an abstract comparison based on their previous experience with ToBI transcription, not based on any exemplar tokens.

Comparisons consisted of (a) pitch accent vs. pitch accent, (b) edge tone vs. edge tone, and (c) pitch accent vs. edge tone. That is, each of the 9 pitch accent possibilities (the 8 accent types plus "unaccented") was compared to each of the 8 other categories. For example, L* vs. H* would be one comparison. Another comparison would be L* vs. unaccented. Another group of comparisons in the questionnaire consisted of comparisons of edge tones with each other. This group included the 3 phrase accent types (L-, H-, and !H) and the 6 combinations of phrase accent and boundary tones (L-L%, L-H%, H-H%, H-L%, !H-L%, and !H-H%). "No edge tone" was included for comparison (making ten types). Each of these 10 types was compared to every other edge tone type. Examples of these comparisons are L- vs. H- , L- vs. L-L% and L-L% vs. L-H%. The final group of comparisons included comparisons of pitch accent types (of which there were 9) with edge tone types (of which there were 10). Comparisons were organized into these three groups to facilitate subjects' judgments.

Subjects judged each pair on a 7-point scale, with 1 being "most similar" and 7 being "most different" (i.e. L*+H H+!H* 1 2 3 4 5 6 7). Each subject made 171 judgments.

2.3. Multi-Dimensional Scaling

The tone similarity judgments for each subject were entered into a matrix, and multi-dimensional scaling was used to create maps of the similarity spaces. SPSS was used for this analysis [6]. MDS uses similarity judgments and represents the similarity among items as distance within a geometrical similarity space. MDS calculates similarity maps in multiple dimensions, and each tone has Cartesian coordinates within the multi-dimensional space. The fit of the map to the data is calculated as the percent of variance accounted for, and this R^2 value is used to select the map with the optimal number of dimensions for the data. In this experiment, the dimensions are expected to be interpretable as corresponding to degree of disjuncture, perceived stress, pitch shape, or pitch level.

3. RESULTS AND DISCUSSION

3.1. Variance Accounted For

The R^2 values for solutions with 1-5 dimensions are given for each subject in Table 1. The R^2 value selected as optimal (that is, showing an improvement of at least .10 from one dimension to the next) is given in bold print for each subject. Subjects differed in the number of dimensions in the optimal solution. Two of the subjects (subject 2 and subject 3) showed the most improvement in R^2 values in increasing from one to two dimensions, with the increase to subsequent dimensions showing little improvement. For these subjects, the two-dimensional map was selected as optimal. The other two subjects (subject 1 and subject 4) showed little improvement in fit beyond the third dimension. For these two subjects, the three-dimensional map was selected as optimal.

	subj.1	subj.2	subj.3	subj.4
1-D	.35714	.50431	.68001	.33923
2-D	.60721	**.73491**	**.83085**	.51747
3-D	**.75238**	.80842	.89181	**.70217**
4-D	.80945	.85304	.92968	.74507
5-D	.84313	.87716	.95236	.82136

Table 1. R^2 values for the 4 subjects for maps increasing from 1-dimensional to 5-dimensional.

3.2. Dimensions

3.2.1. Edge Tones vs. Pitch Accents. For three of the four subjects, the first (and most important, accounting for the most variance) dimension corresponds to the distinction between edge tones and pitch accents. This distinction is depicted in Figure 1, the two dimensional map for subject 3. The line that runs from top to bottom of the graph separates edge tones from pitch accents. All phrase tones and boundary tones are located on the right side of the graph, while all pitch accents are on the left. This pattern is consistent with results from subject 1 and subject 4. The dimensions corresponding to Subject 2's responses are different from all other subjects' in that there was no clear separation of pitch accents and edge tones along any dimension.

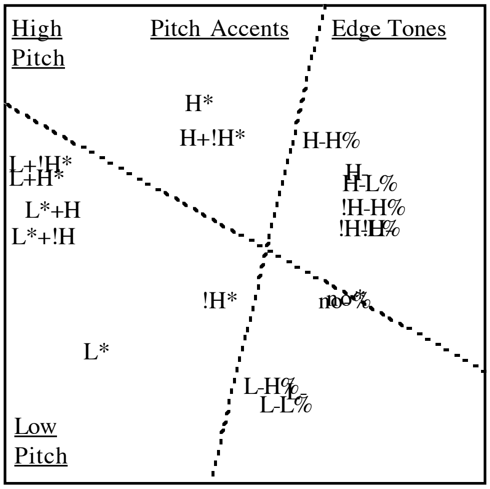

Figure 1. Subject 3's 2-D map.

The separation of pitch accents and edge tones in three of the subjects' results matches the strong intuition that pitch accents and edge tones are different types of pitch events. There are many F0 patterns which might appear identical if considering only the F0 shape. The tonal pattern from a low to high F0 could be analyzable as either a L*+H pitch accent or a L-H% (as discussed in the introduction). However, our results reveal that most transcribers conceptualize these two patterns quite differently. Examples such as this underscore the need for an investigation of the acoustic phonetic features which correlate to this perceptual distinction. This has important implications for automatic labeling of intonation. Other types of acoustic information, which might include intensity and duration in addition to F0, must be incorporated into the automatic labeling algorithm.

3.2.2. Pitch-Related Dimensions. While the first dimension clearly distinguishes between edge tones and pitch accents for three of the four subjects, the subsequent dimensions differ from subject to subject. For Subject 3 (Figure 1), the second dimension is related to pitch level. All edge tones containing an initial H- phrase accent fall in the upper right "quadrant"; two of the three pitch accents containing an initial H component fall in the upper left "quadrant". (The one pitch accent not fitting the pattern, !H*, is usually realized lower than a H* in the pitch range.) Tones beginning with a L component all fall in the lower two "quadrants". The "no-accent" (no*) and "no edge tone" (no-%) both fall on the dividing line between the high and low regions in the map. Thus, there is a progression from lowest pitch level at the bottom to highest pitch level at the top.

Subjects differed in the number of pitch-related dimensions that they have in their results. All subjects have at least one pitch-related dimension, as one would expect given that pitch has been found to be a primary correlate of intonational prominence in English [7]. The second dimension for subject 3 clearly illustrates this, as just discussed. The results from subject 1 also reveal only one pitch-related dimension. However, this subject's pitch-related dimension was the third dimension (with his second dimension being degree of disjuncture).

Subjects 2 and 4 both had two pitch-related dimensions. The two dimensions accounting for subject 2's responses appeared to be pitch related. In general, this subject may have relied on pitch to a greater extent than the others, having had extensive musical training in his life. Because subject 2 only had two dimensions, neither having a split between pitch accents and edge tones, subject 4's responses will be discussed in greater detail. (Results from subjects 1 and 2 will not be discussed further due to space constraints.) Subject 4's first dimension was a distinction between edge tones and pitch accents. Both figure 2, which plots dimension 1 against dimension 2, and Figure 3, which plots dimension 1 against dimension 3, illustrate this clear distinction. Edge tones fall on the right half of each graph, and pitch accents fall on the left. While both the second and the third dimensions are related to pitch, dimension 2 is related to pitch level and dimension 3 is related to pitch movement.

Dimension 2, which defines the y-axis in Figure 2, is related to pitch level. Tones starting at a high level are in the top half of this graph while tones starting at a low level are in the bottom half. The H* pitch accent appears in the upper left quadrant of the graph. This pitch accent begins at a relatively high level in pitch. Directly below is !H*, produced at a lower pitch level. Below that are the various rising accents (all of which start low), the H+!H* (which falls to mid level on the accented syllable), and L*. A similar pattern appears with the edge tones. Edge tones containing H- are in the upper right quadrant, those containing !H- are in the middle, and those containing L- are in the lower right quadrant (except for L- alone, which falls in the middle of the map).

Dimension 3, which defines the y-axis in Figure 3, is related to pitch movement. Falling tones are in the top half of this graph while rising tones are in the bottom half. This generalization holds for both pitch accents and edge tones. H+!H* represents a fall, and it is towards the top of the map. L+H*, which has a rise, is at the bottom. There is a progression from falling pitch accents at the top to rising pitch accents at the bottom of the graph. Similarly, the edge tone often produced with the steepest fall (when realized after the more common pitch accents ending with a H tone, e.g. H*L-L%) appears at the highest level in the upper right quadrant of Figure 3; at the very lowest level is the edge tone, H-H%, realized as a steep rise after a pitch accent ending with a L tone (e.g. L*H-H%). There is a progression from falling pitch patterns to rising pitch patterns between L-L% and H-H% in this dimension.

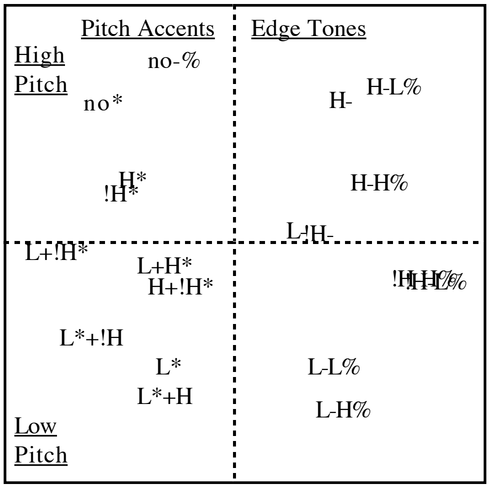

Figure 2. Dimension 1 plotted against dimension 2 for subject 4.

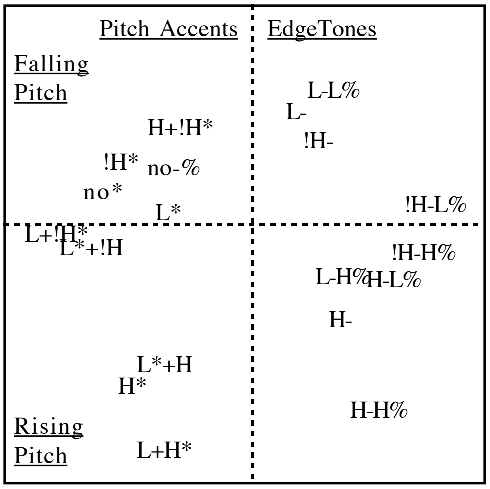

Figure 3. Dimension 1 plotted against dimension 3 for subject 4.

3.3 Tone Similarity

The distance between tones within the maps derived from MDS represents their degree of similarity (or dissimilarity). If two or more pitch events are clustered together, then one can conclude that these pitch events are perceptually similar to one another. This is a way to test intuitions about the similarity between tones. Pitrelli et al. [2], for example, merged L+H* and H* into one category. They also merged each downstepped pitch accent with its non-downstepped counterpart in one of their calculations of intertranscriber reliability.

In order to illustrate the clustering of tones, we will compare the results from subject 3 and 4 (for ease of exposition, having already described their tone similarity spaces above). In comparing these results, it is important to compare clustering along similar dimensions. The first dimension for both subjects was the separation between edge tones and pitch accents and the second dimension was the separation between high and low pitch level (see Figures 1 and 2). Four pitch accents (L+!H*, L+H*, L*+H, L*+!H), all rising from L to H, cluster together for subject 3. This result can be compared to a similar cluster in subject 4's map. For this subject, three pitch accents (L*+H, L*, L*+!H), all containing a L* element, cluster together. L*+!H and L*+H are the common elements shared between these two clusters. Notice, however, that for subject 3, L* is not a part of the group in question while for subject 4, this pitch accent is the common element of all the tones in the cluster. Furthermore, the defining characteristic for subject 3 is a rise within the pitch accent which is not a defining characteristic for any one group within subject 4's data.

3.4 Weighting Tone Similarity

These differences in how tones cluster together could be used to guide the comparisons between labelers' transcriptions. For each subject, tones could be assigned a dissimilarity weight between "0" and "1", where 0 means "the same" (that is, in exactly the same location), and 1 means "completely different" (that is, maximally separated on the map). For each disagreement between transcriptions, the dissimilarity weight between those two transcriptions could be calculated for both subjects and then the overall dissimilarity would be the average of these two weights. In the simplest case, there would be very low dissimilarity between L*+H and L*+!H for both subject 3 and subject 4 since they cluster together for both subjects. So, if subject 3 describes a pitch event as L*+H and subject 4 describes the same pitch event as L*+!H, the average dissimilarity between these transcriptions would also be low. In a more complex example, the dissimilarity rating between L* and L*+H for subject 3 might be closer to 0.5, while it would be closer to 0 for subject 4. Then, if subject 3 transcribed a L* for a pitch event for which subject 4 transcribed L*+H, then the average dissimilarity rating between these transcriptions would thus be around 0.25. Thus, rather than just counting "identity" between tonal transcriptions, this method would allow us to weight dissimilarities using the transcribers' perceptions of tone similarity.

4. CONCLUSION

In conclusion, we have found that intonational transcribers have different perceptual dimensions organizing their tonal similarity spaces. Transcribers were found to have either two or three-dimensional maps. The dimensions were interpreted as edge tones vs. pitch accents, pitch level, and pitch movement.

Within the multi-dimensional maps, clusters of tones were noted. Tones which group together are subjectively similar. Interestingly, tone clusters varied from subject to subject, even when comparing along the same dimensions. This has implications for inter-transcriber reliability tests, since it allows us to go beyond simply counting identical transcriptions as agreements, and since it also allows for a systematic way of merging tonal categories as has been done (on the basis of intuitions and experience with labeling) in previous research. [2, 3, 4] Using the method of weighting disagreements among transcribers gives an overall dissimilarity rating between transcriptions which is more informative. This calls for future research in which clusters of tones and weighted disagreements are calculated in more detail.

ACKNOWLEDGMENTS

This project was funded by ATT. Thanks to Ann Syrdal for her generous support and for her guidance. We gratefully acknowledge the help of our four transcribers: Ken Bame, Craig Hilts, Laurie Maynell, and Julia Tevis McGory. We would also like to thank Keith Johnson, Mary Beckman, and Robert Allen Fox for many helpful discussions. We also appreciate the technical support given by Matt Makashay.

REFERENCES

[1] Beckman, M. E. and G. A. Elam. 1997. Guidelines for ToBI labeling, version 3.0. The Ohio State University.
[2] Pitrelli, J.F., Beckman, M.E., and Hirschberg, J. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. *Proceedings of the 1994 International Conference on Spoken Language Processing*. Yokohama, Japan.
[3] Nakatani, C. 1997 Discourse structural constraints on accent in narrative. In Van Santen, J.P.H., Sproat, R.W., Olive, Joseph, P., and Hirschberg, J. (eds.) *Progress in Speech Synthesis*. Springer: New York.
[4] Jannedy, S. 1997. Acquisition of Narrow Focus Prosody. In Sorace, A., Heycock, C., and Shillcock, R. (eds.) *Proceedings of the GALA '97 Conference on Language Acquisition*. Edinburgh.
[5] Herman, R. 1998. *Intonation and Discourse Structure in English: Phonological and Phonetic Markers of Local and Global Discourse Structure*. PhD Dissertation. The Ohio State University.
[6] SPSS Base System User's Guide. 1990. SPSS Inc.
[7] Beckman, M. 1986. *Stress and Non-Stress Accent*. Foris, Dordrecht.