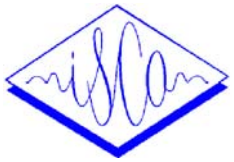

Speech Prosody 2002
Aix-en-Provence, France
April 11-13, 2002

Text-to-tune Alignment in Speech and Song

Karen M. Arnold & Peter W. Jusczyk

Department of Psychology
The Johns Hopkins University, Baltimore, MD
karnold@jhu.edu

Abstract

Musical and linguistic systems are similar in many aspects, including a common timecourse of acquisition and the use of many of the same cues to express structure. Any direct parallels between the systems would be beneficial to infants, who are learning both simultaneously. Using ToBI (Tones and Breaks Indices) analyses of spoken renditions of nursery rhymes and modified ToBI analyses (coding only for locations of high and low targets) of sung versions of the same texts, a strong correspondence was found between melodic and intonational contours of the same materials. These results bear on questions of learnability and bootstrapping in both realms.

1. Introduction

There are many underlying similarities between music and language. Beyond the obvious fact that they both use the auditory modality to convey socially relevant information among conspecifics, the two systems are postulated to have a common evolutionary genesis [1]. In addition, at the most general level, they present the learner with similar problems. Both are hierarchically organized, combinatorial systems [2], where structures are formed from components whose temporal durations and internal complexity can vary [3]. Both have a syntax of sorts, a set of abstract rules or constraints that determine structural well-formedness at a global level. Both also have a finer-grained set of phonotactic rules or constraints governing the combination of individual sound units. Acquisition of the structural systems of music and of language take place concurrently, presumably using many of the same learning processes. To the extent that the parallel progress in both realms is based on similar strategies, one might expect to find specific correspondences in the course of acquisition of music and language.

1.1. Concurrent learning trajectories

A great deal of research has been conducted on exactly when infants gain mastery of various aspects of language or music. Sensitivity to the internal constituent structure of linguistic utterances has been shown at the age of 6 months [4][5]. These infants preferred sentences that were interrupted by pauses artificially inserted at clause boundaries over those that were broken at arbitrary points in the utterance. These results support the claim that by 6 months, infants have already mastered some prosodic cues to syntactic structure, and may have already begun the bootstrapping process.

Likewise, at the same age, infants show sensitivity to the boundaries of musical phrases [6], which lie at the structural level analogous to the clause in speech. Six-month-olds preferred music that had been segmented at phrase boundaries to music that had been interrupted by a pause inserted mid-phrase. These concurrent and closely related findings point to a general ability that is being manifested in two separate but related domains. More specifically, the ability to analyze and decompose larger units based on statistically predictive cues would aid in the discovery of any system's compositional structure and the relations between the smaller constituents.

Music and language both use pitch contour, intensity, and timbre to mark, and to express relationships between, constituents. Despite the clear differences between the two at an abstract level (music lacks a formal semantics, but has an added layer of complexity in its stricter regulation of timing), the question remains as to whether there might be any direct parallels between the **instantiations** of these abstract systems.

1.2. Acoustic cues in common

Analyses of the musical materials used to test the 6-month-olds [7] provide some tempting direct analogies between the cues used to mark structure in music and language. A consistent drop in pitch level was observed before musical phrase boundaries, much like the decrement in pitch range and fundamental frequency commonly found utterance-finally in speech. In addition, and perhaps more convincingly, there was a consistent lengthening of note value (i.e., duration) before musical phrase boundaries. This is the identical acoustic cue used in the same way to signal phrase boundaries in speech [8][9][10]. This cue is even exaggerated in infant-directed (ID) speech [11], providing a potential inroad for infants in the process of discovering the sound patterns of their language.

It would certainly be beneficial for infants, who are learning both systems concurrently, to be able to take advantage of any similarities in the problem spaces of music and language. A logical place to look for these similarities to emerge would be where there is complete overlap: sung text. For example, in a tonal language, where fundamental frequency carries lexical meaning, is there a tendency to construct tunes such that the melodic contours mirror the tonal ones that would be produced if the text were spoken? An investigation of Cantonese popular music [12] found that in these through-composed pieces (where the lyrics and music are created at the same time), there was in fact a high degree of correlation between relative heights of musical pitches and the lexical tone of the words set to them. These results raise the following question: In languages where pitch does not make lexical distinctions but does carry meaning, is there the same tendency to retain the relative pitch heights within sentences' intonational contours in the melodies to which they are set?

The current study was designed to address this issue, using English as the medium. The prediction was that if a "good" melodic contour is in any way constrained to match an intonational one, the phenomenon should be observable when sung and spoken versions of the same text are compared.

2. Methods

Putting aside the controversy surrounding the sufficiency of infant-directed (ID) speech to drive the bootstrapping process, there are clearly characteristics that **could** be helpful to a beginning learner. ID speech is more attractive to infants [13], commanding longer spans of greater attention. In addition, it has been characterized as providing strong, salient, and consistent prosodic cues to syntactic structure [9][11][14][15]. This information would then be available to the infant for deeper processing. The phenomenon of simpler ID material promoting processing is actually observed in the musical realm. Melodies showing characteristics of ID music (e.g., stylized contours roughly analogous to those found in ID speech, use of principle scale degrees) support more detailed analysis of their constituents and the relationships between them [16][17]. Since the transparency and magnitude of the cues seem to be enhanced in ID material [18], it is reasonable to assume that any direct correspondence between the cues used would be most visible in that realm. Accordingly, we focused on music and texts that were both clearly intended for infants: nursery rhymes.

2.1. Materials

Three classic nursery rhymes were selected for analysis, based on three criteria. The materials were: 1) reasonably familiar, 2) able to be read as prose, i.e., without overly strong impositions of meter, and 3) set to melodies that were unfamiliar to the speakers who produced the stimuli, to avoid bleed-over effects from the melodic contour. They contained a wide range of syntactic and pragmatic constructions, including questions, quotations, and relative clauses. The texts used were as follows:

- **May Day Carol:** I've been wand'ring all the night, and the best part of the day. Now I'm returning home again. I bring you a branch of May. [19]
- **Simple Simon:** Simple Simon met a pieman going to the fair. Says Simple Simon to the pieman: "Let me taste your ware." Says the man to Simple Simon: "Do you mean to pay?" Says Simon: "Yes, of course I do!" and then he ran away. [20]
- **Mistress Mary:** Mistress Mary, quite contrary, how does your garden grow? With cockle shells and silver bells, and fair maids all in a row. [21]

The texts were recorded a total of five times each by female native speakers of American English. Two speakers, naïve to the purpose of the study, were presented with the above materials typed as connected prose rather than verse broken into rhyming couplets. They produced two recordings of each text, one in ID speech (elicited by instructing the speakers to act as if they were telling the story to a very young child) and one in the speech register used for conversing with adults. In all cases, the speakers were instructed to read the texts as prose, rather than metered verse. A fifth recording of each of the songs was produced by a semi-professional vocalist. They were sung *a cappella*, with no vibrato, on the pitches and in the meter dictated by the sheet music.

2.2. Transcriptions

The speech samples were digitized, and transcribed using the Tones and Breaks Indices (ToBI) system of analysis, roughly following Pierrehumbert & Hirschberg's description [22]. ToBI is a descriptive and theoretical system that codes for several aspects of prosody, including locations and types of pitch prominences and degree of disjuncture between words. In this study, the aspect that was most important was the location and type of pitch accents (relative f0 targets). The *xwaves*™ program and suite of macros and scripts were used, and the transcriptions followed the conventions set forth in Guidelines for ToBI Labelling [23].

Approximately 10% of each of the recordings was also transcribed by another experienced ToBI labeller, and the one discrepancy found was discussed and resolved in a collaborative fashion. (NB: The discrepancy was a minor one, between a H* and a !H* near the end of an utterance.)

It has been shown that infants are attuned to the relationships among pitch levels (i.e., their relative height and location in a contour) rather than their absolute pitch [24]. In addition, the study of Cantonese song and lexical tone described earlier [12] found that a fruitful way to analyze notes in a melody when comparing them with prosodic phenomena was as relative pitch targets, not absolute pitch levels. This meshes well conceptually with the way that ToBI tags f0 contours, so, in order to accommodate the differences between speech and song but still retain some degree of comparability between the analyses, the following modifications to the ToBI system were used when labeling the song samples:

- Only high (H*) and low (L*) pitch targets were labeled, based on maxima and minima of the song's f0 track that were aligned with stressed text syllables.
- Some minor pitch movements (of 2 semitones or less) were disregarded as derived from musical stylistic demands rather than actual pitch targets.
- High and low targets were coded relative to key changes, in much the same way as in standard ToBI they are coded with respect to pitch range compressions and resets.

Two points are worth mentioning: First, the simpler analyses of the musical samples were completed before the speech samples were tagged using the full ToBI system, with a significant time lapse between the two types of tagging. This was done to ensure that the coding of the music was not influenced by the labeller's knowledge of the details of the coded speech samples. Second, the labeller who performed the second analysis of the speech had neither heard the sung renditions of the texts nor seen the tagged samples, so there was no possible contamination of their verification coding.

2.3. Analyses

Following Syrdal & McGory's [25] intertranscriber reliability measures, within each text, on a word-by-word basis, the placement of pitch accents was tabulated across all renditions. For example, all five renditions placed a H* pitch accent on the first word of *Mistress Mary*. On the second word, three spoken renditions placed a downstepped H*, one used a L*, and the sung rendition used a H*. In contrast, on the sixth and seventh words, none of the renditions realized a pitch accent. For the purpose of comparison between the results from the two different coding schemata, all pitch accents that were primarily high (i.e., H*, !H*, and L+H*) were collapsed to a single category of "high target." Likewise, L* and L*+H were considered to be "low targets." For simplicity, the simplified categories will be henceforth referred to as H* and L*. No boundary tones were included in the analyses, because they were not coded for in the music samples. Five

measures of consistency among renditions were then calculated for each word of each text (see following explanations), and means for each measure were computed for each text. When calculating these means, only data from words where at least one rendition assigned it a pitch accent were included.

2.3.1. Presence or absence of pitch accent

To investigate the consistency of the locations of pitch accents without regard for type, the following calculation was performed, first among the speech samples alone, and then including the sung sample. The general question addressed by this measure was whether pitch accents occurred in the same **locations** among renditions.

$$\frac{\text{Number of renditions containing any pitch accent}}{\text{Total number of renditions}} \quad (1)$$

2.3.2. Agreement on pitch accent type

Then, to evaluate the consistency of pitch accent types, the following calculation was performed, again on both speech samples alone and then including the song. This measure determined whether, in locations where pitch accents occurred, the **type** of pitch accent agreed among renditions.

$$\frac{\text{Largest number of renditions agreeing on accent type}}{\text{Number of renditions giving that word a pitch accent}} \quad (2)$$

2.3.3. Agreement with majority

The final measure was whether on a given word, the pitch accent type of the sung version agreed with the **majority** of the spoken renditions that assigned that same word a pitch accent. Each word was assigned a binary value of 1 (agree) or 0 (doesn't agree). This metric addresses the question of whether the contour of the musical tune broadly corresponds, at least in the location and type of high and low targets, to that of a **common** spoken one. If the spoken renditions were evenly split as to pitch accent type and the song matched one or the other attested type, it was counted as agreeing. If either the song received a pitch accent and none of the spoken versions did, or vice versa, it was counted as non-agreeing.

2.3.4. Expected values

In order to compare the correspondence values computed by the above equations with random assignment of pitch accents, expected values were calculated for each of the five measures. The following prior probabilities were assumed:

- It is equally likely that a word be said with or without a pitch accent. P(pitch accent on word x) = .50
- Given that a word receives a pitch accent, the two types of pitch accent are equally likely. P(H*) = P(L*) = .50

The expected values were then used in performing an unpaired *t*-test for each song on each measure.

3. Results and discussion

The results reported below are mean correspondence values for each text, averaged across all words that received a pitch accent in at least one rendition, as compared with expected values, using unpaired *t*-tests calculated for each song. The final column represents the significance level of the result of each individual test.

Table 1: Presence/absence of pitch accent, within speech samples only

Song	EV	Mean	*p*
Simple Simon	.50	.73	<.05
May Day Carol	.50	.83	<.05
Mistress Mary	.50	.95	<.05

Table 2: Presence/absence of pitch accent, speech and song samples combined

Song	EV	Mean	*p*
Simple Simon	.50	.69	<.05
May Day Carol	.50	.68	<.05
Mistress Mary	.50	.98	<.05

It is clear from the results contained in Tables 1 and 2 that there is a significant degree of agreement in **where** pitch accents are realized. Also, 2-tailed *t*-tests comparing results from speech only and speech and song combined showed no significant differences between the samples (all *p*s >.05). From Tables 3 and 4, however, we can see that there is much less agreement on pitch accent **type**, both among the spoken versions of a text and among spoken and sung versions of that text. None of the *t*-tests even approached significance.

Table 3: Agreement on pitch accent type, within speech samples only

Song	EV	Mean	*p*
Simple Simon	.81	.88	>.05
May Day Carol	.81	.89	>.05
Mistress Mary	.81	.85	>.05

Table 4: Agreement on pitch accent type, speech and song samples combined

Song	EV	Mean	*p*
Simple Simon	.79	.91	>.05
May Day Carol	.79	.90	>.05
Mistress Mary	.79	.87	>.05

The question remains, however, whether the contours of a musical melody might correspond to a **typical**, or at least to a **possible** intonational contour, rather than to all, or a specific set of, attested contours. The final analysis addressed this question by comparing the pitch accent types found in the songs to the pitch accent types in the majority of the spoken versions of the same text. Table 5 demonstrates that pitch targets realized in sung text do indeed correspond in type with those often realized in like locations in spoken text.

Table 5: Agreement on pitch accent type, song vs. majority of speech

Song	EV	Mean	*p*
Simple Simon	.30	.43	<.05
May Day Carol	.30	.56	<.05
Mistress Mary	.30	.93	<.05

These results bring up a potential new area of inquiry with respect to the utility of the infant-directed register, both in speech and song. The question of whether infants make use

of the cues said to characterize the register is predicated on the claims that the cues are present, consistent enough to be helpful, salient enough to be detected, and only then put to work in the task of bootstrapping. To the extent that some of these cues may be present in multiple domains, both the salience and the utility could be boosted, and this correspondence would provide the perfect vehicle for parallel learning. When investigating the input used in the task of language acquisition, it may not be enough to look at infant-directed speech alone, or even speech alone; music might be another important piece of the puzzle for infants.

4. Conclusions

This study presents a new methodology for comparing spoken and sung pitch contours in an objective manner, where they can be subjected to statistical analyses rather than descriptions or subjective impressions. It opens the way for many avenues of investigation. One might expand the range of musical styles and genres included in the analysis, as well as the number and diversity of speakers. The variability among ID and AD renditions of a text should be compared, as an assessment of the consistency of the cues postulated to be paralleled in music. The temporal dynamics of actual ID song (i.e., that produced in the presence of infants) is also an area that bears further investigation, to see whether any modifications to the strict meter of the song might correspond to speech phenomena like phrase final lengthening. In short, this is just a first step towards understanding how the cues used in bootstrapping might also be present in other domains.

5. Acknowledgments

This paper is dedicated to the memory of Peter Jusczyk, whose support was instrumental in the early phases of this project.

Many thanks go to Joseph Kisenwether for his help with the statistical analyses reported here.

This research was supported by a scholarship to the Ohio State University Department of Linguistics summer workshop series "Spoken Language in Context: Methods and Models" to KMA, and by a Research Grant from NICHD (15795) and a Senior Scientist Award from NIMH (01490) to PWJ.

6. References

[1] Brown, S., 2000. The musilanguage model of music evolution. In *The Origins of Music*, N. Wallin, B. Merker, & S. Brown (eds.). Cambridge, MA: MIT Press, 271-300.

[2] Lerdahl, F.; Jackendoff, R., 1983. *A generative theory of tonal music*. Cambridge, MA: MIT Press.

[3] Deutsch, D.; Feroe, J., 1981. The internal representation of pitch sequences in tonal music. *Psychological Review* 88, 503-522.

[4] Hirsh-Pasek, K; Kemler-Nelson, D.; Jusczyk, P.; Wright-Cassidy, K.; Druss, B.; Kennedy, L., 1987. Clauses are perceptual units for young infants. *Cognition* 26, 269-286.

[5] Nazzi, T.; Kemler Nelson, D.; Jusczyk, P.; Jusczyk, A.M., 2000. Six-month-olds' detection of clauses embedded in continuous speech: Effects of prosodic well-formedness. *Infancy* 1(1), 123-147.

[6] Krumhansl, C.; Jusczyk, P., 1990. Infants' perception of phrase structure in music. *Psychological Science* 1(1), 70-73.

[7] Jusczyk, P.; Krumhansl, C., 1993. Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *Journal of Experimental Psychology: Human Perception and Performance* 19(3), 627-640.

[8] Klatt, D., 1975. Vowel lengthening is syntactically determined in connected discourse. *Journal of Phonetics* 3, 129-140.

[9] Lederer, A.; Kelly, M., 1992. Prosodic information for syntactic structure in parental speech. Manuscript, Department of Psychology, University of Pennsylvania.

[10] Scott, D., 1982. Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America* 71, 996-1007.

[11] Fisher, C.; Tokura, H., 1996a. Acoustic cues to grammatical structure in infant-directed speech: cross-linguistic evidence. *Child Development* 67, 3192-3218

[12] Chan, M., 1987. Tone and melody in Cantonese. *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society.* Berkeley: BLS, 26-37

[13] Fernald, A., 1985. Four-month-olds infants prefer to listen to motherese. *Infant Behavior and Development* 8, 181-195.

[14] Fernald, A.; Simon, T., 1984. Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology* 20, 104-113.

[15] Fernald, A.; Taeschner, T.; Dunn, J.; Papousek, M.; de Boysson-Bardies, B.; Fukui, I., 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language* 16, 477-501.

[16] Cohen, A.; Thorpe, L.; Trehub, S., 1987. Infants' perception of musical relations in short transposed tone sequences. *Canadian Journal of Psychology* 41, 33-47.

[17] Trehub, S.; Thorpe, L.; Trainor, L., 1990. Infants' perception of good and bad melodies. *Psychomusicology* 9, 5-19.

[18] Trehub, S.; Trainor, L.; Unyk, A., 1993. Music and speech processing in the first year of life. *Advances in Child Development and Behavior* 24, 1-35.

[19] Bacon, D. (arr.), 1993. May Day carol. *24 Favorite Nursery and Folk Songs*. Wellesley, MA: Kodaly Center of America, 25.

[20] Elliott, J., 1828. Simple Simon. *Mother Goose's Nursery Rhymes and Nursery Songs.* Springfield: McLoughlin Bros., 31.

[21] Elliott, J., 1828. Mistress Mary. *Mother Goose's Nursery Rhymes and Nursery Songs.* Springfield: McLoughlin Bros., 1.

[22] Pierrehumbert, J.; Hirschberg, J., 1990. The meaning of intonation contours in the interpretation of discourse. In *Plans and Intentions in Communications*, P. Cohen, J. Morgan, & M. Pollack (eds.). Cambridge, MA: MIT Press, 271-312.

[23] Beckman, M.; Elam, G., 1997. Guidelines for ToBI labeling, version 3.0. The Ohio State University Research Foundation.

[24] Trehub, S.; Bull, D.; Thorpe, L., 1984. Infants' perception of melodies: The role of melodic contour. *Child Development* 55, 821-830.

[25] Syrdal, A.; McGory, J., in press. Inter-transcriber reliability of ToBI prosodic labeling.