An Empirical Method for Comparing Pitch Patterns in Spoken and Musical Melodies: A Comment on J.G.S. Pearl's "Eavesdropping with a Master: Leos Janáček and the Music of Speech"

ANIRUDDH D. PATEL
The Neurosciences Institute, San Diego

ABSTRACT: Music and speech both feature structured melodic patterns, yet these patterns are rarely compared using empirical methods. One reason for this has been a lack of tools which allow quantitative comparisons of spoken and musical pitch sequences. Recently, a new model of speech intonation perception has been proposed based on principles of pitch perception in speech. The "prosogram" model converts a sentence's fundamental frequency contour into a sequence of discrete tones and glides. This sequence is meant to represent a listener's perception of pitch in connected speech. This article briefly describes the prosogram and suggests a few ways in which it can be used to compare the structure of spoken and musical melodies.

Submitted 2006 June 1; accepted 2006 June 6.

KEYWORDS: *speech intonation, musical melody, perception, melodic contour, computational models*

WHILE the term "melody" is commonly used to refer to pitch patterns in both speech and music, it is clear that spoken and musical melodies have salient differences in terms of structure and perception. For example, it is common to find oneself humming a catchy musical tune, while one rarely notices or explicitly remembers spoken pitch patterns. This is because musical melodies are designed to be aesthetically interesting objects, while pitch patterns in ordinary speech are not (cf. Patel, 2003). Instead, pitch variations in speech serve communicative functions such as distinguishing lexical items (in tone languages), making prosodic boundaries, indicating the pragmatic status of words and utterances, signaling a speaker's affective state, and so forth. Spoken pitch patterns perform these functions without themselves becoming a focus of attention for ordinary listeners.

Not all listeners are ordinary, however. Over the centuries, a number of musicians have found themselves captivated by spoken melodies and have invested considerable effort in treating these melodies in a musical framework. The work of Joshua Steele (1700-1791) is perhaps the best known in this regard. Steele wanted to capture the prosody of talented actors and orators, so that "the types of modern elocution may be transmitted to posterity as accurately as we have received the musical compositions of Corelli" (quoted in Kassler, 2005). To this end, he modified the musical staff to allow the transcription of spoken pitches down to a quarter-tone (i.e., one-half of a semitone), and developed a way to indicate the direction and extent of pitch glides within syllables (see Figure 2 of Kassler, 2005, for an example).

Leos Janáček's (1854-1928) musical transcriptions of intonation are perhaps less well known, at least in linguistic circles. Janáček began transcribing intonation contours of Czech speech around the end of the 19th century, and continued for over thirty years. Unlike Steele, Janáček focused on the prosody of ordinary speech, often doing his transcription surreptitiously so that the speaker would not be influenced by knowing his or her speech was being recorded (Christiansen, 2004). According to Christiansen (2004), Janáček's interest in speech melodies stemmed from a belief that intonation patterns revealed the emotional and psychological state of a speaker, independent of the semantic content of the words being spoken.

The scholarly literature on Janáček's transcriptions has been rather sparse, and the article by Pearl (2006) in this journal will help bring Janáček's work to a wider audience. For researchers interested in the relationship of music and language, one obvious question about Janáček's transcriptions is their relationship to phonetic reality. Pearl (2006) suggests that Janáček's 'tunelets of speech' "were intended as

accurate-as-possible descriptions of his momentary perceptions." Yet Janáček used Western music notation to transcribe intonation contours, meaning that he accommodated his descriptions to a pre-established grid of pitches and intervals. Since spoken pitch patterns do not conform to a fixed system of intervals, it would be interesting to know how Janáček's transcriptions relate to phonetic reality. Unfortunately, we do not have recordings of the utterances that Janáček transcribed. Even if we had such recordings, however, Pearl (2006) argues that a deeper problem is that "the melodies of speech are not self-evident, objective features of the world. There is no established heuristic powerful enough to extract all and only the pitches that a hearer will perceive in any stream of speech…"

Very recently, a quantitative model of speech intonation perception has been advanced which has the potential to overcome this problem, and thus to open the way to empirical comparisons of spoken and musical melody. The remainder of this essay briefly describes the "prosogram" model and suggests a few ways in which it might be used to examine music-language relations. (For a more detailed description of the prosogram, and one application to speech-music research, see Patel, Iversen, & Rosenberg, 2006).

The central notion behind the prosogram is that the raw fundamental frequency (Fo) contour of a sentence, although an accurate physical description of the speech signal, is not the most accurate representation of intonation as it perceived by human listeners. In particular, empirical research suggests that pitch perception in speech is subject to several perceptual transformations. One of these is perceptual segregation of the Fo contour into syllable-sized units due to the rapid spectral and amplitude fluctuations in the speech signal. A second is temporal integration of Fo within the syllable. That is, if the Fo contour within a syllable is short and has relatively little change, then listeners perceive a single pitch which is a time-weighted average of the intra-syllabic Fo movement. If the amount of intrasyllabic Fo change is above the "glissando threshold", then a glide (or multiple glides) is perceived. The prosogram instantiates these transformations via an algorithm which takes an audio file as an input and produces a "tonal score" as an output, converting a sentence's original Fo contour into a sequence of discrete tonal segments. An example of the model's output is given in Figure 1.

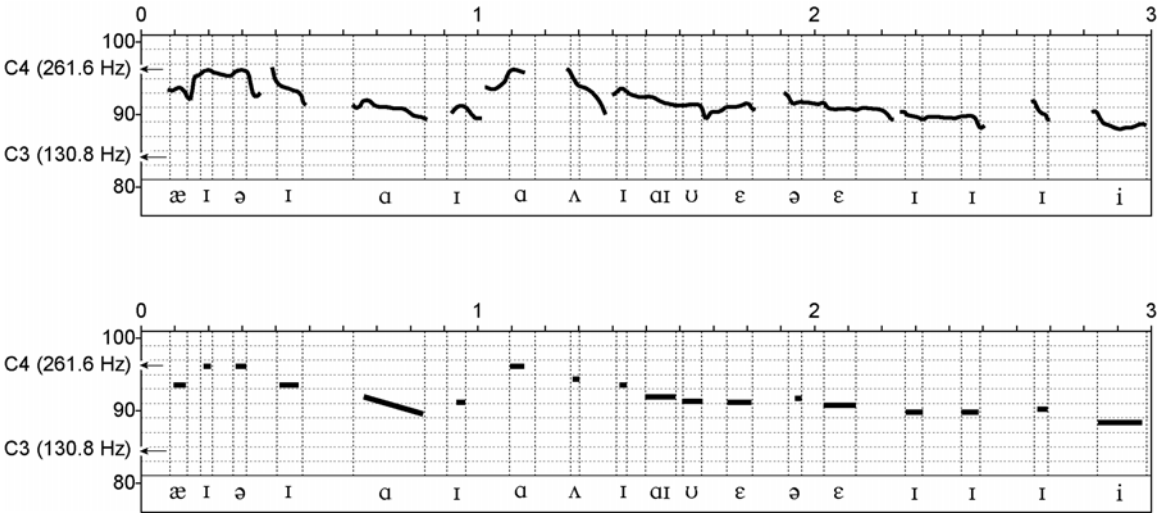

Fig 1. Illustration of the prosogram, using the British English Sentence "Having a big car is not something I would recommend in this city" as uttered by a female speaker. (To download an audio file of the original recording, click this link: https://kb.osu.edu/dspace/bitstream/1811/24011/1/EMR000009b-pateltest.ogg). In both graphs, the horizontal axis along the top shows time in seconds, the vertical axis shows semitones relative to 1 Hz (thus 90 st corresponds to 181.02 Hz, see Note [2]). Horizontal lines on the grid show 2-semitone increments, and arrows mark the frequencies corresponding to musical C3 and C4 for reference. Characters along the bottom of each graph are the International Phonetic Alphabet (IPA) symbols for the vowels in this sentence. The temporal onset and offset of each vowel is indicated by vertical dashed lines above that vowels' IPA symbol. The upper graph shows the original fundamental frequency (Fo) contour of the sentence, while the lower graph shows the prosogram. In this case, the prosogram was based on the

sentence's vowels, and required the onset and offset time of each vowel as input. (Different phonological units, such as syllables or syllable rimes, could also be used as input. A more recent version of the prosogram does automatic segmentation of speech into syllable-like units based on patterns of voicing and loudness.) In this example, the prosogram has assigned level tones to all vowels save for the vowel in "car", which was assigned a glide. The prosogram is freely available from: http://bach.arts.kuleuven.be/pmertens/prosogram/.

Figure 1 reveals why the prosogram is useful to those interested in comparing speech and music. The perceptual representation of intonation produced by the prosogram is quite music-like, consisting mostly of level pitches. These pitches are displayed on a semitone axis, in accordance with the logarithmic nature of pitch perception in speech (Nolan, 2003) [1]. An important point about the prosogram is that the pitches it assigns to syllables are not constrained to follow any particular musical scale. Thus the pitches of the prosogram can "fall between the cracks" of the Western chromatic scale, and the intervals between pitches can have non-integer values (e.g., 4.2 semitones).

The fact that the prosogram converts a sentence's Fo contour into a sequence of discrete pitches opens the way to quantitative comparisons of spoken and musical melody. One such study, inspired by Janáček's work, would be to compare a musician's transcription of speech intonation to a prosogram analysis of the same sentences. One could then see how much "warping" of spoken pitch values must take place in order to fit speech intonation into Western musical notation. For example, in 2004 the jazz musician Rudresh Mahanthappa released a CD called "Mother Tongue" containing instrumental pieces inspired by aural transcriptions of speech intonation in a variety of Indian languages (http://pirecordings.com/pi14/). As part of this project, Mahanthappa created detailed musical transcriptions of short spoken monologues from a number of different speakers. These transcriptions could be compared to prosogram analyses of the same recordings in order to study how speech melody is perceived by a keen musical ear.

Another use for the prosogram would be to compare the patterning of spoken and musical melody in song. A number of studies have examined tone languages to see how the tonal patterns of the spoken lyrics of a song relate to the melodic contour to which these lyrics are set (e.g., Richard, 1972; Wong & Diehl, 2002). These studies have relied on the fact that there are categorical phonological distinctions between linguistic lexical tones, so that each syllable of the lyrics can be assigned a phonological tone height (e.g., low, mid, high, etc.). The contour formed by this "lexical tone melody" can then be compared to the contour formed by a song's musical melody, to study the degree of concordance between them. Using the prosogram, a similar kind of study could be conducted in intonation languages. Specifically, a speaker could read the lyrics of song, and prosograms of the resulting utterances could be made. (Ideally, the speaker would be unfamiliar with the song so that his or her intonation pattern would not be influenced by prior knowledge of the musical melody). The melodic contour of the speech melody could then be compared to that of the musical melody to which the lyrics were set by the composer. Are there sections of the song where the spoken and musical melodic contour are either very similar or very different, and do these relate to interesting points in the musical structure? Does the amount of overall convergence between spoken and musical pitch contours differ in different genres of music (e.g., folk songs vs. art songs)? Are songs with a high degree of convergence between the contours of their spoken and musical melodies easier to learn or remember than songs with a low degree of convergence?

In short, a new tool from phonetics has opened the way to systematic and empirical comparisons of spoken and musical pitch patterns, and to investigating the little-explored region that lies between the linguistic and musicological study of melody. One imagines that Janáček would be pleased.

NOTES

[1] It is important to note that the prosogram was developed on the basis of perceptual research with native speakers and listeners of "intonation languages", i.e., languages which do not use pitch to distinguish lexical items. Hence its relevance to pitch perception in tone languages such as Mandarin is not yet clear.

[2] Formula for converting the from Hz to semitones (s.t.) relative to 1 Hz: s.t. = 12*log2(X), where X is the Hz value. Formula for converting s.t. to Hz: $Hz = 2^{(X/12)}$, where X is s.t. relative to 1 Hz.

ACKNOWLEDGMENTS

I thank Patrick Wong and Bruno Repp for helpful comments. Supported by Neurosciences Research Foundation as part of its research program on music and the brain at The Neurosciences Institute, where ADP is the Esther J. Burnham Fellow.

REFERENCES

d'Alessandro, C. and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language,* Vol. 9, pp. 257-88.

Christiansen, P. (2004). The meaning of speech melody for Leoš Janáček. *Journal of Musicological Research,* Vol. 23, pp. 241-63.

Kassler, J.C. (2005). Representing speech through musical notation. *Journal of Musicological Research,* Vol. 24, pp. 227-39.

Mertens, P. (2004). The Prosogram : Semi-automatic transcription of prosody based on a tonal perception model. *Proceedings of Speech Prosody 2004,* Nara (Japan), pp. 23-6.

Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. In: *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona*, pp. 771-4.

Patel, A.D. (2003). A new approach to the cognitive neuroscience of melody. In: I. Peretz and R. Zatorre (Eds.), *The Cognitive Neuroscience of Music*. Oxford: Oxford Univ. Press, pp. 325-45.

Patel, A.D., Iversen, J.R., & Rosenberg, J.C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *Journal of the Acoustical Society of America,* Vol. 119, pp. 3034-47.

Pearl, J.G.S. (2006). Eavesdropping with a Master: Leoš Janáček and the music of speech. *Empirical Musicology Review,* Vol. 1, No. 3, pp. 131-66.

Richard, P. (1972). A quantitative analysis of the relationship between language tone and melody in a Hausa song. *African Language Studies,* Vol. 13, pp. 137-61.

Wong, P.C.M., & Diehl, R.L. (2002). How can the lyrics of a song in a tone language be understood?. *Psychology of Music*, Vol. 30, pp. 202-9.