Phonological equivalence of pitch level in speech

Paper presented at a conference on Music and Language, Tufts University, July 2008

D. R. Ladd, University of Edinburgh

It's fairly obvious that voice pitch goes up and down in the speaking voice just as it does in the singing voice, and it's tempting to talk about the 'scale' of speech pitch in the same way as we talk about musical scales. However, it appears that the scale metaphor doesn't fit speech pitch very well, or rather, that spoken scales and musical scales are qualitatively and quantitatively different. What I want to do to today is to give a concise introduction to my work on the 'scales' involved in speech pitch, and speculate briefly about why speech and music might be different.

Let me begin by making a few observations about the scientific investigation of pitch. In everyday language pitch is normally treated as one-dimensional. In English, it can go up and down, but not also sideways or round and round in circles, and the English high/low metaphor is widespread in the languages of the world. Other languages use different metaphorical dimensions – sharp/dull, bright/dark, light/heavy, young/old, and many others – but a common feature of the way many or even all languages treat the percept of pitch is that it moves along a single dimension. Similarly, classic experimental work using the methods of psychophysics has shown that the primary physical correlate of pitch is fundamental frequency (F0), which is also a single dimension. This work has also produced classic psychophysical functions relating F0 to perceived pitch, rather like other psychophysical work that relates physical to perceptual dimensions in sensory features like loudness, brightness, and so forth. So everyday language seems to be based on something real that we can investigate in the lab.

But there's a problem. When I say this work has produced psychophysical *functions*, that's exactly what I mean. Depending on the psychophysical questions you ask, you seem to get different functions for the relation between F0 and pitch – Barks and Mels and ERB units and semitones. And when the answer to an empirical question depends on how you phrase the question, one reasonable conclusion is that, however it's phrased, it's the wrong question. At the very least, classical psychophysics isn't going to tell us everything we want to know about pitch.

This is the background to the last 50 years or so of experimental research on pitch in music. For a lot of this work, classical psychophysics is essentially irrelevant, and it's tempting to translate that

into the conclusion that classical psychophysics is actually misleading or downright wrong. Roger Shepard was quite explicit about this (1982: 306):

> Until recently, attempts to bring scientific methods to bear on the perception of musical stimuli have mostly adopted a psychoacoustic approach. The goal has been to determine the dependence of psychological attributes, such as pitch, loudness, and perceived duration, on physical variables of frequency, amplitude, and physical duration ... or on more complex combinations of physical variables ...
>
> By contrast, the cognitive psychological approach looks for structural relations within a set of perceived pitches independently of the correspondence that these structural relations may bear to physical variables. This approach is particularly appropriate when such structural relations reside not in the stimulus but in the perceiver ...

So what are we missing if we look at musical pitch just as psychophysics? The most conspicuously unsatisfactory aspect of psychophysical work on pitch, in addition to the fact that it appears to give us multiple competing answers, is that it doesn't lead us to expect the existence of **octave equivalence**. Octave equivalence seems to be fundamental to all musical scales and may occur in other animals, so it clearly belongs somewhere in our understanding of pitch. In order to allow for octave equivalence, Shepard proposed the now familiar spiral representation of pitch shown in Fig. 1, which separates pitch height (on the vertical dimension) from chroma (on the circular dimension) and gives direct visual representation to octave equivalence. He went on to develop the search for 'geometric approximations' to the structure of musical pitch in order to express more and more complex cognitive relations between pitches – standard musical notions like interval and key and resolution.

However, subsequent work has investigated these cognitive relations without necessarily pursuing the goal of better geometric approximations of pitch. In her 1990 book Carol Krumhansl quite explicitly distinguishes between understanding the cognitive foundations of musical structure and developing a graphic representation of music pitch relations. I read Krumhansl's work not so much as rejecting psychophysical conclusions but as simply saying that we need to take **context** into account if we want to understand the cognitive structuring of pitch in music. What produces the conflicting results in psychophysical experiments is failing to anchor the experiments in context. So for example – this is my example, not hers – it seems to be a matter of psychophysical fact that the bottom A and Bb notes on the piano are harder to tell apart than the A and Bb in the octave that starts on middle C. Nevertheless, that growly low note may be reliably interpreted as A or Bb depending on the harmonic and melodic context. Reduced psychoacoustic distance and constant

structural distance can both be true – the cognitive structuring of musical pitch doesn't necessarily invalidate findings in psychophysics.

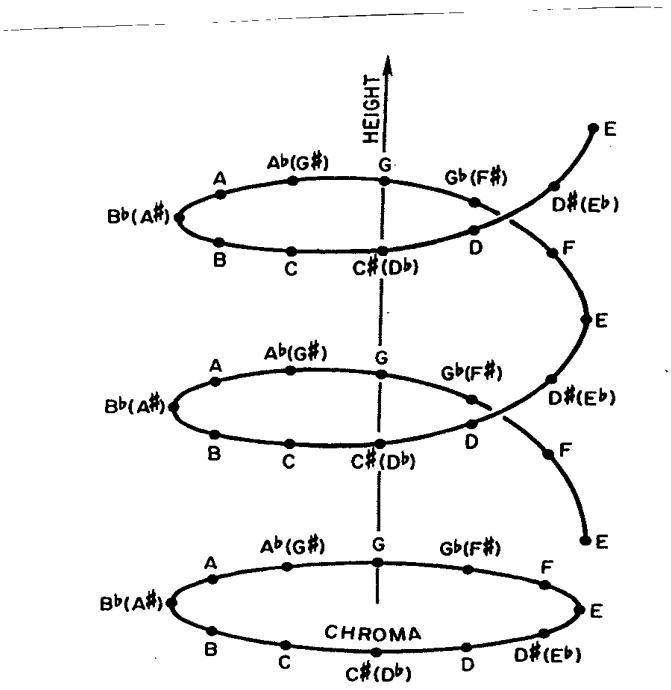

Figure 1. Shepard's spiral representation of pitch height and chroma.

So what I want to do today is to take context seriously, and to compare across the biggest contextual divide of all, namely that between speech and music. Phoneticians, like music psychologists, have come up with conflicting psychophysical findings about 'the' appropriate scale for speech pitch. The variable conclusions are based on a variety of empirical evidence: psychophysics-style experiments like judging the relative prominence of accented words, comparisons of pitch in male and female voices, observations of so-called declination (the tendency of pitch to drop substantially across an utterance), and much more. Here too I think the quest for a single generally valid scale or representation of pitch has blinded us to the importance of cognitive structuring and context – as with musical pitch, these phonetic experiments may provide conflicting results because they implicitly involve different higher-level cognitive contexts.

Phonological equivalence of speech pitch level

What spoken language has that is lacking in music is **phonological equivalence**. This can be illustrated with an example from Jackendoff (1972). Take the sentence *Fred ate the beans*, pronounced in two different dialogue contexts:

(1) A: What about Fred? What did he eat?
B: Fred ate the beans.

(2) A: What about the beans? Who ate them?
B: Fred ate the beans.

There are two clearly distinct ways of pronouncing the B sentence, which signal whether the A sentence it's responding to is a question about Fred or a question about the beans. The difference is signalled by the height of the pitch peaks on the two words, by the depth of the pitch valley in between, and various other phonetic details. Now, the absolute pitch level of those peaks and valleys is obviously going to be different from speaker to speaker, and, within a given speaker, it may differ from one mood or setting to another. But neither of these kinds of differences – differences between speakers' voice ranges, and differences of expressiveness within a given speaker – affects our ability to determine whether the speaker is answering a question about Fred or about the beans. Despite the conspicuous differences in range, we're able to extract some kind of phonological equivalence over all the versions that are about Fred and all the version that are about the beans.

So understanding how phonological equivalence works is going to be a significant piece of the full story of how speech pitch works, and phonological equivalence is still not well understood. An obvious hypothesis would be that it works in terms of **pitch changes or relative pitch differences.** For example, there might be a rise of x semitones on *Fred* and a rise of y semitones on *beans*, or perhaps the difference between the pitch peak on *Fred* and the pitch peak on *beans* is specified as x semitones in one case and as y semitones in another. But it has become very clear that this isn't actually how phonological equivalence works. Instead, it turns out that phonological equivalence seems to be based on very lawful correspondences of **pitch level**. To see this, what we have to do is base our comparisons across speakers or across expressive contexts in the same speaker on specific **target points** in contours.

To see how this works, assume that the peaks and valleys in any given version of *Fred ate the beans* are **targets** – pitch levels that the speaker is aiming at. These peaks and valleys are shown in idealized form in Figure 2, which shows two different renditions of sentence (1)B (the sentence that is appropriate in response to a question about what Fred ate). The two lines show the contours that we might expect if the sentence was spoken by two different speakers with different ranges, a wide or expressive range (the red line) and a narrow or monotonous range (the blue line). There are five 'targets': two peaks, the medial valley, and the starting and ending pitches.

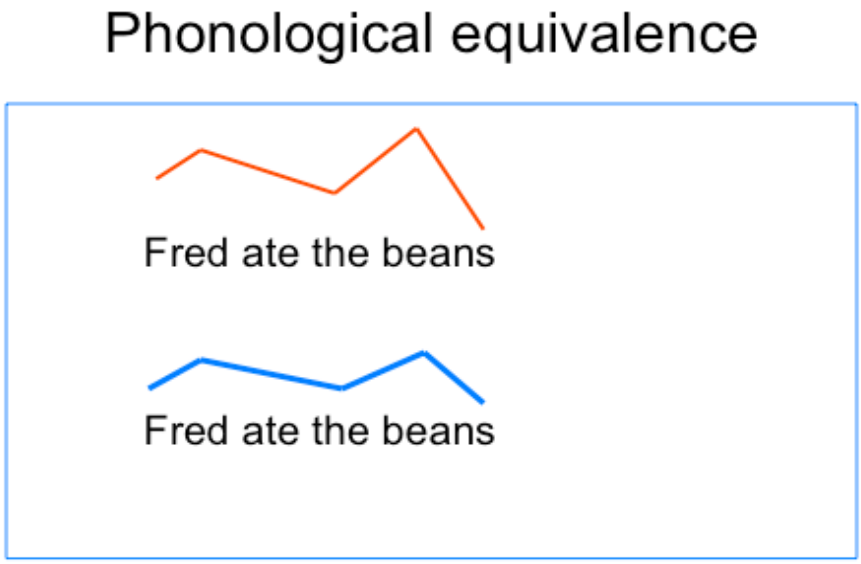

Figure 2: Idealized representations of the intonation contour in sentence (1)B, as spoken by two different speakers, one with a wide range (red) and one with a narrow range (blue). The three local minima and the two local maxima can be considered 'targets' for purposes of comparing across speakers.

If we plot the value of a set of such target point across two speakers (like the red speaker and the blue speaker in Figure 2), or across different degrees of emphasis or expressiveness in the same speaker, we find that we obtain consistently high correlations. Figure 3 shows a graph of one speaker against another for a corpus of material read aloud, while Figure 4 shows, for two different individuals, graphs of a speaker's normal and raised voice plotted against each other. It can be seen that speakers are doing something quantitatively very precise when their voice pitch goes up and down as they speak.

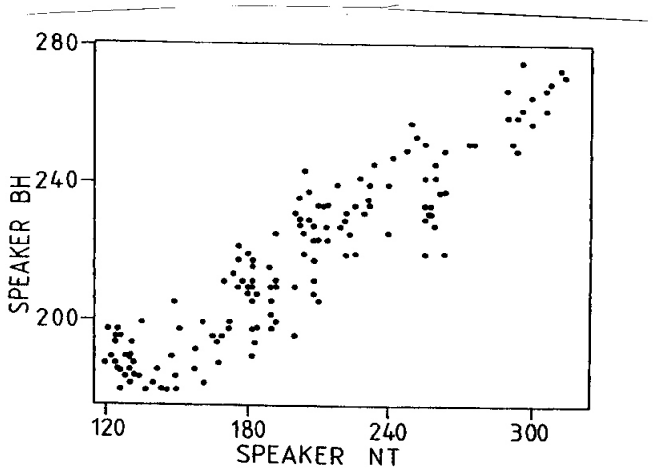

Figure 3. Inter-individual correlations of pitch level. Based on Danish read speech data from Nina Grønnum, using carefully controlled speech materials where it was possible to identify "the same" point in a given utterance spoken by two different speakers; each point on the graph plots the mean F0 value of a given point in Speaker NT's speech against the same point in speaker BH's speech. Axes show F0 in Hz.

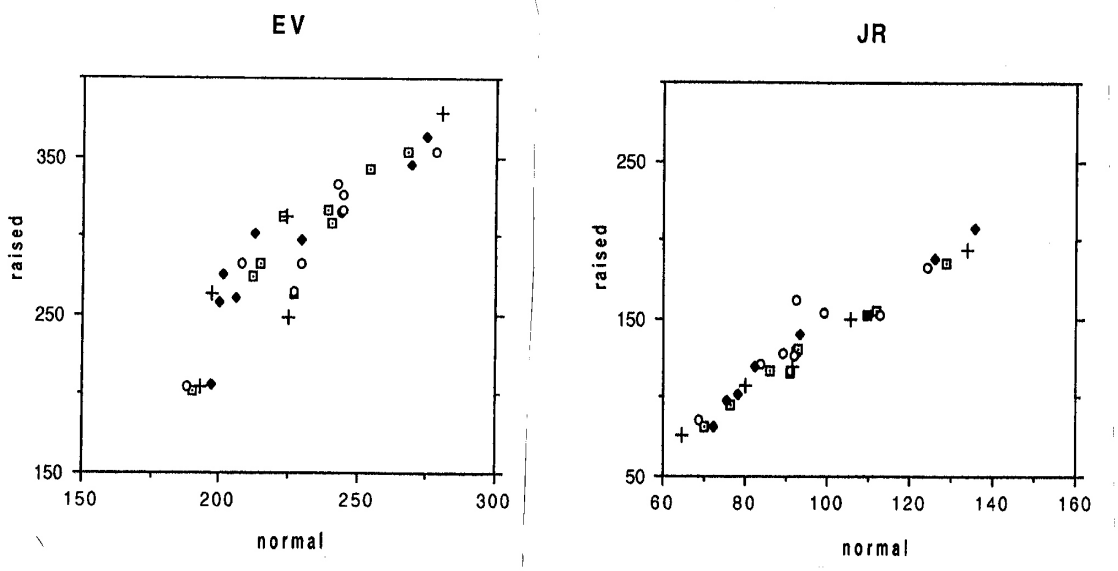

Figure 4: Within-individual correlations of pitch level. Based on a corpus of controlled read speech in Dutch collected by Jacques Terken and myself, in which speakers read a series of sentences and then in a later part of the recording session read the same sentences but raising their voice. Similarly to Figure 2, each point on the graph represents the F0 value of a point in an utterance as spoken in normal voiced plotted against the same point spoken in raised voice.

However, these correlations don't imply anything about pitch intervals in the musical sense. If we have one speaker with a monotonous voice (like the blue speaker in Figure 2) and another speaker with a lively voice (like the red speaker in Figure 2), it's clear that the lively speaker's pitch changes and pitch intervals are going to be wider than the monotonous speaker's. But what the existence of the correlation tells us is that, **within their respective pitch ranges**, the lively speaker's targets are going to be spaced out proportionally in the same way as the monotonous speaker's. This is shown in Figure 5. There is a kind of speaker specific pitch scale, but it **doesn't** involve any kind of constant musical intervals.

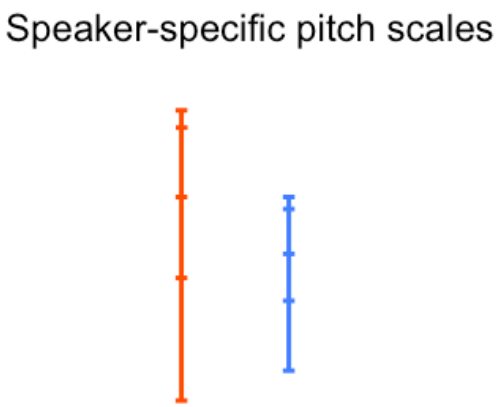

Figure 5. Speaker-specific pitch scales, based on the idealized contours in Figure 2. For each of the two speakers, the top target on the scale represents the pitch peak on beans and the second-highest target represents that on Fred. The three lower targets represent, from high to low, the sentence-initial pitch, the sentence-medial minimum, and the final low.

Moreover, what is really interesting, and possibly somewhat unexpected, is that you get a more or less identical story if you look at tone languages (languages like Chinese or Thai or Yoruba, in which the pitch patterns function like the consonants and vowels to tell one word from another). From the point of view of a speaker of English these languages are weird and mysterious, and it's tempting to think that they might be processing pitch in some fundamentally different way. For example, Diana Deutsch has speculated that tone languages must make use of something like absolute pitch (Deutsch et al. 1999). But this hypothesis is unnecessary. What emerges clearly from phonetic studies of tone languages is that they actually manage phonological equivalence in exactly the same way we do in languages like English. That is, there are consistent quantitative

correlations between one speaker and another in the way the pitch level distinctions are realized in speech. But the actual pitch **intervals** used by different speakers are not equal in a musical sense, any more than the interval between *Fred* and *beans* is equal for different speakers. Instead, phonological equivalence is based on **some kind of proportional relation to the speaker's pitch span.**

This can be illustrated with data from Mambila, a language from the Nigeria-Cameroon borderland spoken by about 100,000 people. (Mambila recordings were made available to me by Bruce Connell). Mambila has four level tones, numbered from 1 (highest) to 4 (lowest). Three of these are illustrated in the following minimal set (the combination of the highest level with [ba] doesn't happen to mean anything):

(3) a. [bo4 ba2 mo4] 'my bags'
b. [bo4 ba3 mo4] 'my palms'
c. [bo4 ba4 mo4] 'my wings'

Table 1 shows the average measured values in Hz for the four pitch levels, as spoken in contexts like the one in example (3), for five different Mambila speakers; Figure 6 shows the same data in graphic form in a way comparable to the hypothetical English data shown in Figure 5.

Mambila tone distinctions

mean pitch levels for 5 speakers

	CD	SM	BJ	VM	MD
High (1)	166	120	136	252	200
High-Mid (2)	142	109	120	214	176
Low-Mid (3)	130	102	115	197	167
Low (4)	117	92	104	172	149

[data from Bruce Connell]

Table 1: Average pitch values (in Hz) of four Mambila tones, for five speakers.

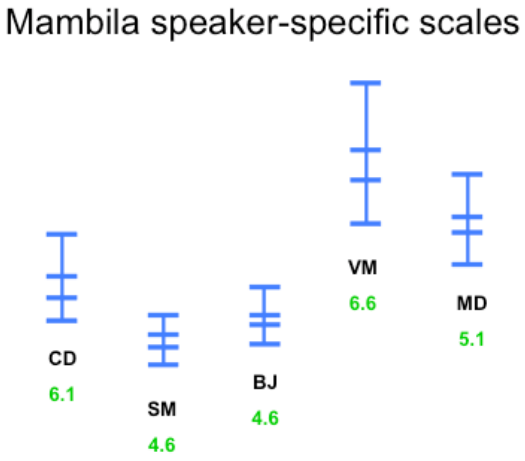

Figure 6: Data from Table 1 represented graphically; compare Figure 5. The number in green below each scale is the width of the range in semitones.

The key empirical finding exemplified by the Mambila data (and based on a wide variety of other data discussed in more detail in Ladd 2008, chapter 5) is as follows: phonological equivalence is based on **constant proportional level** within a given pitch span – 22% or 85% or whatever it happens to be – and not on constant pitch intervals expressed on any known musical or psychophysical scale. And what that in turn means is that absolute pitch is irrelevant to phonological equivalence; phonological equivalence depends on the way a given pitch relates to the phonetic context defined by the speaker's pitch span. This recalls Krumhansl's emphasis on context, and it originally seemed to me that this context-dependence is reminiscent of structural relations within a key in music. But several people have suggested that it's more comparable to the kind of normalisation that we all have to do when we adapt to the vowel space of speakers with different accents and/or with vocal tracts of different size. In any case, it's clear that speech pitch doesn't involve any analogue to musical chroma – or any analogue to octave equivalence.

Conclusion

So why the difference between music and speech? I think the main reasons are functional. In speech, pitch serves two communicative purposes that are essentially independent of one another. First, there's phonological equivalence, which I've just been illustrating. But second (and almost certainly prior from an evolutionary point of view) is the paralinguistic or expressive modification of pitch span – widening or narrowing the range of pitches to sound happy or angry or sad or bored or whatever. We can't judge phonological equivalence relative to some fixed speaker-specific

context, but only to a **current** pitch span that we have to calculate on the fly on the basis of the sequence of pitches we hear. (How we do this isn't clear, and this is one of the reasons I think phonological pitch equivalence may involve some similarities with music cognition and not just with phonetic normalisation, but that is a topic for another talk.) In music, on the other hand, unlike speech, we ***must*** pay attention to chroma and octave equivalence. When men and women or adults and children sing together, they necessarily sing in different ranges, but paying attention to octave equivalence guarantees that the consonances and dissonances will be preserved. Without that, we have the equivalent of a group reciting a prayer or a pledge – the ups and down coincide roughly, so there is phonological equivalence, but there is no consonance and no music.

So what I'm suggesting is that things like phonological equivalence and harmonic relations are comparable organisational principles that are specific to their respective domains. And I want to reiterate the point I attributed to Krumhansl earlier: that these higher-level cognitive relations have no necessary bearing on the psychophysics of pitch and should not be part of the attempt to understand speech pitch 'scales' or the geometry of musical pitch. I think that one reasonable way of looking at this whole cluster of questions would be to say that the psychophysical foundations of pitch in speech and music are the same, but that the cognitive organisation that we impose on pitch is quite variable. We know that it is variable from one musical tradition to another; I suggest that there is an even more fundamental difference between speech on the one hand and music – any musical tradition – on the other. But if we assume that the psychophysical foundations of speech pitch and musical pitch are not essentially different, then the difference between these two systems must lie at a higher level.

References

Deutsch, Diana; Henthorn, Trevor; Dolson, Mark (1999). Absolute pitch is demonstrated in speakers of tone languages. *Journal of the Acoustical Society of America* 106: 2267.

Jackendoff, Ray (1972). *Semantic interpretation in generative grammar.* MIT Press.

Krumhansl, Carol L. (1990). *Cognitive foundations of musical pitch.* Oxford University Press.

Ladd, D. Robert (2008). *Intonational phonology*, 2nd ed. Cambridge University Press.

Shepard, Roger N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review* 89: 305-333.