## Native wails

November 6, 2009 @ 8:51 am · Filed by Mark Liberman under Prosody, Psychology of language, The language of science

« previous post | next post »

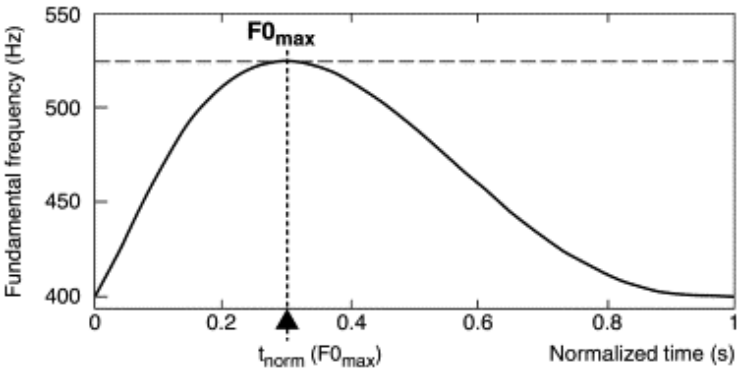
In today's newspapers and magazines:

> "Newborns cry in their native language".
> "Babies cry with an accent within the first week of life".
> "Babies cry wiith the same 'prosody' or melody used in their native language by the second day of life".
> "Newborn babies mimic the intonation of their native tongue when they cry".
> "French babies cry in French, German babies cry in German and, no doubt, the wail of an English infant betrays the distinct tones of a soon-to-be English speaker".

The science behind these statements is in a paper released yesterday: Birgit Mampe, Angela D. Friederici, Anne Christophe and Kathleen Wermke, "Newborns' Cry Melody Is Shaped by Their Native Language ", *Current Biology*, in press. Does it support these journalistic generalizations? Before reading the paper, I give ten-to-one odds against, on the general principle that journalistic statements involving generic plurals are almost never true. *Mesdames et messieurs, faites vos jeux*. Let's spin the wheel.

Mampe et al. recorded and analyzed the crying of 30 French babies (11 female, 19 male; mean age 3.1 days, range 2–5 days) and 30 German babies (15 female, 15 male; mean age 3.8 days, range 3–5 days). They recorded 2500 cries, of which they selected 1254 "simple cries containing single rising-and-then-falling melody arcs" — an average of 21 per French baby, range 3–54; an average of 18 per German baby, range 10–38.

Cries of the type selected had pitch contours that the authors schematize like this:




They found the location of the pitch peak in each cry, and expressed it in terms of normalized time, where the overall duration of the cry is set to 1; and the same analysis was performed for amplitude.

It was by no means true that French babies cried in one consistent way, and German babies in another. As the experimenters write:

> Only simple cries containing single rising-and-then-falling melody arcs were analyzed. These cry types were selected because they predominate in the crying of healthy newborns. These melody arcs can be assigned to four basic melody types: (1) quickly rising and slowly decreasing melody: left-accentuated type—"falling contour"; (2) slowly rising and quickly decreasing melody: right-accentuated type—"rising

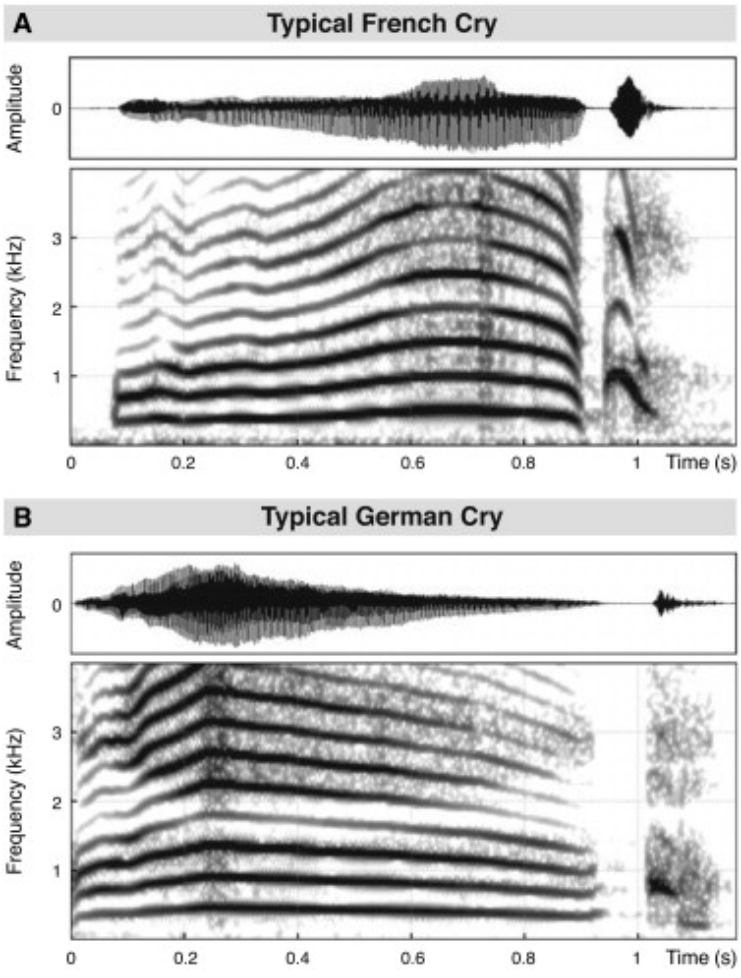
> contour"; (3) symmetrical rising-and-then-falling melody: symmetric type; and (4) a relatively stable fundamental frequency with a rising or falling trend: plateau type. […]
>
> Newborns of both groups generated all four basic melody types typical at that age. This observation reflects a general aptitude for generating melodies with varying contours and explains the observed partial data overlap in Figure 1.

They don't actually give the proportions of French and German cries in their four categories, but they do quantify their results as follows:

> [A] marked difference in the median values of $t_{norm}$(F0max) points to group-specific preferences for produced melody contours (French group, 0.60 s; German group, 0.45 s). The arithmetic means of $t_{norm}$(F0max) were significantly different in French (0.58 ± 0.13 s) and German (0.44 ± 0.15 s) newborns (Mann-Whitney test, $p < 0.0001$). Whereas French newborns preferred to produce rising melody contours, German newborns more often produced falling contours. These results show a tendency for infants to utter melody contours similar to those perceived prenatally. A significant difference was also found for the intensity maxima of melody arcs [ $t_{norm}$(Imax): mean 0.59 ± 0.12 versus 0.47 ± 0.12 for French group versus German group; Mann-Whitney test, $p < 0.001$].

Regular readers of Language Log, and other sensible people, will now be estimating the effect size — and this is a fairly large one. A difference of 0.58-0.44 = 0.14 in time-normalized F0 peak location, with a pooled standard deviation of 0.14, is an effect size of d = 1.0. That is, the mean values for the two groups are separated by one standard deviation. And the same is true for the difference in mean amplitude peak locations.

Despite this large and impressive difference, the authors indulge in a bit of rhetorical exaggeration by presenting the following picture of "typical" French and German cries:




The F0 peak of the "typical" German cry is at about 0.25, seconds in a vocalization whose expiration phase lasts about 0.95 seconds, for a normalized peak time of about 0.26. This is more than a standard deviation below the reported German mean value of 0.44, and thus is hardly "typical". The F0 peak of the French cry is at

0.68 seconds, in  a cry whose expiration lasts about 0.9 seconds, for a normalized time of about 0.75. Again, this is more than a standard deviation greater than the French average of 0.58.

This technique of cherry-picking atypical "typical" values for rhetorical effect is fairly common in scientific writing, but it should alert us to the fact that the authors are perhaps not trying quite as hard as they should to disprove their own hypothesis.

Though I'd bet that their findings are basically valid, there are a few reasons to wonder whether equally large differences would be found in a replication.

For one thing, the babies were recorded in different places, apparently under different circumstances, by different experimenters. Specifically, the French babies were recorded in the Cochin Hospital in Paris. The authors write that "For the German newborns, existing digital sound files of cries that were recorded as part of the German Language Development Study (http://glad-study.cbs.mpg.de) were used", but they don't say whether these were also hospital recordings or were recorded at home. (So it's possible that the headlines should have read "Babies in hospital cry differently from babies at home"…) In any case, the German recordings were apparently made by a different set of people.

Given the procedures described in the paper, there are also several ways that unconscious experimenter bias might creep into the analysis phase. The process of choosing roughly half of the recorded cries as suitable for analysis is one place that this might happen. In doing this sort of selection, there are always a lot of marginal cases, where the decision as to whether an observation fits the criterion (here, whether the cry is a "single rising-and-then-falling melody arc") could go either way. These marginal cases can be influenced by the experimenter's perception that a particular exemplar is "typical" or "not typical" of expected patterns.

And another place where experimenter expectations could affect the data is in marking the time-pints of amplitude and pitch maxima. This was apparently done by hand on narrow-band spectrograms, and again was apparently not done "blind" — that is, the experimenters knew whether they were coding French babies or German babies. And again, there would have been many cases with flat-topped peaks, or with multiple peaks, where choosing the location of the maximum point would have been a somewhat ambiguous task.

Finally, the choice about where the cry begins and (especially ends) is often ambiguous, and the national difference in mean F0-peak location was only about 140 milliseconds.

As a matter of intellectual hygiene, it's a good idea either to make all such decisions "blind" to the independent variables of interest, or else to use automated techniques of measurement with parameters chosen before the data is analyzed. (It's a good idea to publish the data as well, but that's another story.)

So let's sum up. This is a really interesting and suggestive study, which needs to be replicated to be entirely convincing. It finds a fairly large difference in the distribution of pitch and amplitude profiles of French and German neonates, with the French babies tending to produce cries with later peaks that the German babies. The effect size in the reported data is a large one (d=1.0), which is large enough that (if the estimates generalize to new data) a randomly selected French or German baby would be correctly classified as French or German, on the basis of one cry, about 2/3 of the time.

The authors attribute this difference to the typical differences between French and German intonation patterns, through exposure in the womb. It's certainly true that the proportion of final rises in French speech is much greater than in German. But French non-terminal intonational patterns — the ones that generally involve final rises — are not at all like the pitch contour of the "typical" French neonate shown in this paper. The adult phrases will typically involve a large rise on the first or second syllable — to a peak that will often be the highest point in the phrase — with a subsequent fall and then a rise at the very end of the very last syllable, so that there is no final fall. (Today's breakfast hour is over for me, but I promise to give more information on this question within the next few days.)

So if the differences in this experiment's data are caused by pre-natal experience with adult intonations, the explanation must be a somewhat indirect one. And it's not clear to me why the authors reject without explicit consideration the hypothesis that the babies were responding to a few days of French vs. German "motherese".

Oh, and the journalistic generalizations were false as an expression of the authors' findings. Of course.

November 6, 2009 @ 8:51 am · Filed by Mark Liberman under Prosody, Psychology of language, The language of science

Permalink

---

## 25 Comments

1. **Chris said,**

   November 6, 2009 @ 9:44 am

   *This technique of cherry-picking atypical "typical" values for rhetorical effect is*

   I would have completed this sentence "intellectually dishonest". Contrasting that with the way you completed it is a rather sad comment on scientific publishing, especially if this piece has already passed peer review without any of the reviewers finding this worthy of comment.

   Your methodological concerns seem fairly major, too. I wonder why they didn't occur to the experimenters?

2. **Mark P said,**

   November 6, 2009 @ 9:51 am

   Again and again your discussions show a point that journalists don't, won't or can't understand: experiments don't yield facts, they yield experimental results. Of course it doesn't help when the researchers forget it or overanalyze the results for the sake of a satisfying conclusion.

3. **Daniel Ezra Johnson said,**

   November 6, 2009 @ 10:01 am

   30 babies in each group isn't bad, but how can we discuss the statistical results without separating between-baby and within-baby variation? Put another way, are their reported p-values way too low? Unless each country's babies are quite similar, I suspect so. I may keep making the same point, but I think it's a good one. More at http://www.ling.upenn.edu/~johnson4/johnson_panel.pdf.

4. **Steve Silberman said,**

   November 6, 2009 @ 10:03 am

   Mark: I appreciate the close analysis, but I have to ask you: Do you hope to improve the accuracy of science journalism by making glib, snide, overbroad statements about journalists that are the mirror image of the glib, overbroad conclusions you accuse journalists of drawing from studies like this? Or is there some other goal? As a science journalist myself (who didn't write about this study), I'm thoroughly aware of the maddening shortcomings of a lot of science journalism. But I'm not sure what statements like the "ten-to-one" statement accomplish here, beyond creating the sort of self-satisfied effects that also plague bad science journalism.

5. **Benjamin Zimmer said,**

   November 6, 2009 @ 10:24 am

   More on this from John Wells here.

6. **mgh said,**

   November 6, 2009 @ 10:33 am

   the 'typical' case would not necessarily be one near the mean of the distribution, but rather near its mode.

   (I don't know if this would make any difference in Mark's critique.)

7. **uberVU - social comments said,**

   November 6, 2009 @ 11:07 am

   **Social comments and analytics for this post…**

   This post was mentioned on Twitter by interests: Language Log: Native wails http://bit.ly/4pUmgB…

8. **Don Monroe said,**

   November 6, 2009 @ 11:13 am

   Since they have the data, it seems to me that they should do the whole analysis over, blinded. The opportunity for biased data selection could create the whole effect, not just enhance it.

   [(myl) I should say that I don't know for sure that they didn't do the selection and annotation "blind", just that they don't say anything one way or the other about the issue, and that usually means that blind techniques weren't used.]

9. **Sili said,**

   November 6, 2009 @ 11:21 am

   > prenatally

   So the parents and everyone else around them kept schtumm for two-three days?

   As you say, there's really no excuse for not depositing original data in this age of electronic publishing. I've been out of research for some years now, but if I get to go back, I hope I'll remember to put all spectra and the like up as supplementary material.

   I see Zimmer has already linked the *rien ne va plus*.

10. **seth edenbaum said,**

    November 6, 2009 @ 12:29 pm

    "Do you hope to improve the accuracy of science journalism by making glib, snide, overbroad statements about journalists that are the mirror image of the glib, overbroad conclusions you accuse journalists of drawing from studies like this?"

    Ah… the pull of narrative. Pleasure and pattern.
    Giving meaning to the world.

11. **Jan van Santen said,**

    November 6, 2009 @ 12:53 pm

    Even if the cases "not suitable for analysis" were excluded without any bias at all (e.g., some automated measure would find exactly the same group means of 0.58 and 0.44), inclusion of these cases with their more-than-likely poorly defined pitch and amplitude peaks may very well increase the within-group

standard deviations and hence decrease the effect size. This is a general problem with Cohen's d [ =(mn1-mn2)/sd] definition of effect size: even when there is no obvious bias that could affect the means there still can be a bias that artificially decreases the within-group standard deviations and hence increases the effect size. Unprincipled elimination of outliers, even if this does not affect the group means, is another good example. Usage of nonparametric measures of effect size such as ROC area can reduce — but not eliminate — this type of bias.

12. **Terry Collmann said,**

    November 6, 2009 @ 1:09 pm

    "Do you hope to improve the accuracy of science journalism by making glib, snide, overbroad statements about journalists that are the mirror image of the glib, overbroad conclusions you accuse journalists of drawing from studies like this?"

    Yes, I think he does.

    But given the pressures on journalists to produce stories that grab attention, rather than reflect accurately what is being said, I'd give you much longer than 10 to 1 against him succeeding.

    (Oh, and I'm a journalist …)

    [(myl) And it doesn't help that scientific PR often encourages the "stories that grab attention".

    I recognize that there are many excellent science journalists, and that science stories are often written by journalists on other beats. But my main goal in writing posts like this, aside from finding a way to tell a story about a scientific paper that's in the news, is to encourage readers to adopt a suitably skeptical attitude when they encounter generic statements about groups.]

13. **Theo Vosse said,**

    November 6, 2009 @ 1:53 pm

    Surprisingly there is no information about the cry length distribution. Since they squashed all cries onto the same length, a difference of 30% in length could cause the effect as well. Or a few good outliers.

    A more serious problem is of course the classification. How do you distinguish a falling or a rising cry from a rise+fall? Where is the threshold? You could very well argue that only fall means F0 at 0 seconds, and only rise means F0 at end of cry (or 1 second when following their boxing method). If the data doesn't have a problem, this would only increase the effect size. So why didn't they analyze it this way, or report the difference in rising/falling preference?

    But if the goal is to show influence of the environment, then what can a rise-fall contour prove if the most prominent difference between the languages is a final rise, especially when German babies seemingly prefer falling and French rising cries? Wouldn't that be enough? This has the hallmarks of a fishing expedition (i.e., get loads of data, classify them arbitrarily, and then report on the class with the largest effect as if it were your original hypothesis). I know the publication pressure at Max Planck institutes. And it was published in "Current Biology", which is not known to me for its linguistic orientation, so I guess it has been rejected elsewhere.

14. **Ellen said,**

    November 6, 2009 @ 2:22 pm

    The first headline listed, "Newborns cry in their native language", at least has the merit of being so obviously false as to make it clear that it's not anywhere close to literally true.

15. **D.O. said,**

November 6, 2009 @ 7:47 pm

I am a bit surprised that the authors use units (s) for the normalized time. Makes no sense. They definitely should publish the original data at least if they are not planning to do more analysis on it. Then the authors quite naturally have dibs. If NIH grant was used in the process, publishing original data is almost a requirement, if I understand it correctly. Otherwise, the effect is large (not merely statistically significant) and if some "Occam's broom" was in use, well, we shall see.

[(myl) The effect as reported is very large, and the phenomenon is in principle very interesting, whatever it may mean. So I expect that others will try similar things, as it's not hard to do. Meanwhile, it would be nice to be able to see their data.

There are good independent reason to create and publish a large archive of neonate cries, I think, since there are known clinical signs that it would be nice to be able to screen automatically, and there might be other useful diagnostic information that is not now known.]

16. **Gracie Vindicated « Buttle's World said,**

November 6, 2009 @ 8:03 pm

[…] shocking turn of events it seems that the popular press has completely misrepresented the science. And the science wasn't so hot to begin with. This technique of cherry-picking atypical "typical" values for rhetorical effect is […]

17. **Interesting Stuff: Early November 2009 « The Outer Hoard said,**

November 7, 2009 @ 12:06 am

[…] Claims that French and German babies cry differently have been met with scepticism. […]

18. **v said,**

November 7, 2009 @ 3:40 pm

Ouch, this post was just slashdotted. Well it was linked in the highest modded comment to the article about this so…

19. **Crying Babies [The Frontal Cortex] » iThinkEducation.net! said,**

November 7, 2009 @ 4:02 pm

[…] Important qualifications from the always lucid Language Log: This is a really interesting and suggestive study, which needs to be replicated to be entirely […]

20. **matt said,**

November 10, 2009 @ 1:31 pm

I agree with you generally about science journalism, but this seems to be a case where the scientists, as you suggest, did most of the work of overinterpreting. The difference between the paper title "Newborns' Cry Melody Is Shaped by Their Native Language" and the headline "Newborns cry in their native language" is almost non-existent!

21. **Leonardo Boiko said,**

November 13, 2009 @ 12:06 pm

That's quite a bit of peer review you guys got going here. Somewhere, sometime, the researchers will read this thread, and I kind of feel bad for them…

[(myl) You shouldn't feel bad for them. They've found suggestive evidence for a really interesting phenomenon. Precisely because the phenomenon is so interesting (i.e. so unexpected against the background of certain widely-shared assumptions, and at the same time so well aligned with various alternative competing assumptions), their evidence deserves to be examined critically, and I expect that it will be, and not just here.

If the finding holds up under scrutiny and under replication, then it's a really important one, and the paper will be widely cited for a long time, along with the research that it engenders. If the effect vanishes under scrutiny and replication, then the paper will mostly be forgotten. And perhaps the most likely outcome is something in between, where perhaps the effect would replicate but in a more complicated way or to a smaller degree.

I can't pretend to know how it will turn out. But the study's authors should be commended for trying an interesting experiment and for reporting the results in a clear way.]

22. **A Little Science about Language and Babies « 500 Words on Words said,**

November 17, 2009 @ 6:10 pm

[…] just in case you're still curious, here's a fairly in-depth analysis of that same study from the Language Log blog at University of […]

23. **Jim Scobbie said,**

November 18, 2009 @ 2:31 pm

Mark says that "This is a really interesting and suggestive study, which needs to be replicated to be entirely convincing." However, I find it hard to imagine getting funding to undertake an explicitly corroborative study, or that it would be easy to get it published in a (top) journal if it were undertaken and the results matched. The replication of research studies is extremely important, and not just for training the next generation of researchers (ideally at masters level, though perhaps not in this case) in current techniques. Yet because the pressure is on always to creatively extend or challenge, not to replicate, instead of results that might straightforwardly confirm or disconfirm a previous study, we typically get a set of partial replications which do neither, but instead appear to provide an interesting or intriguing twist on the groundbreaking original. I would like to see more kudos for good, careful, convincing replications in addition to the deserved focus on speculative, creative, innovative and conceptually inspiring studies.

[(myl) I agree; but a well-designed follow-up study could simultaneously attempt a replication of this work, and also include data and modeling work that would extend it.]

24. **Skrik « Mellom turrfisken og veden said,**

November 20, 2009 @ 1:47 pm

[…] er ikkje den fyrste som har blogga om dette, men til liks med Mark Liberman på Language Log la også eg fort merke til figurane som viser eit «Typical French Cry» og eit «Typical German […]

25. **Can you tell the language of the mother from her baby's cry? « Nicolas Claidière said,**

February 21, 2010 @ 11:27 am

[…] are closer to the melody of their mother's tongue than to that other tongues (but see Mark Liberman Language Log post for some methodological […]