How similar are pitch contours derived from 'imaginary' student interactions to those derived from real interactions?

[1]Monja Knoll and [1, 2]Lisa Scharrer

[1]Department of Psychology, University of Portsmouth, Portsmouth, UK
[2] Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Germany

monja.knoll@port.ac.uk

Abstract

This study evaluated the use of imagined interactions in speech research, by comparing speech addressed to imaginary speech partners with natural speech addressed to genuine interaction partners. The shape of pitch contours derived from target words of samples of imaginary infant (IDS), foreigner (FDS) and British adult (ADS) directed speech produced by ten female students was compared to that derived from an existing data set of natural IDS, FDS and ADS. For the pitch contour shape analysis we used a standard qualitative approach and a previously evaluated novel algorithmic method. We found no significant difference in pitch contour shape between IDS, ADS and FDS in the imaginary interactions. Unlike our previous findings for natural speech where IDS had been characterised by exaggerated contours compared to both adult conditions, all three speech types had a similar distribution of pitch contour shape. This contrast between the present findings and the interactions with genuine speech partners, suggests that speech obtained from imaginary interactions should be used with caution.

1. Introduction

Infant-directed speech (IDS) is acoustically and phonetically modified compared with adult-directed speech (ADS) [e.g. 1, 2]. The well recognised acoustic modifications of IDS include exaggerated pitch contours, increased mean pitch, hyperarticulation and high emotional affect [e.g. 3, 4, 5]. These IDS modifications probably have a linguistic role in language acquisition, but they may also have emotional-attentional functions [e.g. 1, 3]. Separation of these functions by comparing IDS with other linguistic (foreign-directed speech: FDS) [6] or emotional affective groups (e.g. pets or partners) [3, 5] has become a main goal in IDS-centred research. Several methodologies have been used in these investigations, with IDS having been compared to ADS using students and imaginary scenarios in the laboratory [7, 8], and in natural interactions in the mother's home [5, 6].

Advantages and limitations are inherent in each of these methodologies. Close control of the environment (e.g. background noise levels, room acoustics and consistency of recordings) is a key feature of laboratory-based studies. The use of students imagining talking to relevant groups in the laboratory is also more convenient and time efficient than using genuine interaction partners in a natural setting. In the case of IDS and FDS, finding an infant or foreign confederate who is available over an extended duration is often problematic. Conversely, laboratory interactions may be contrived and unnatural. For instance, previous exposure to specific listener groups and ability to 'act' can affect how well an individual can perform when imagining talking to those listener groups. The comparability of the findings of these studies to genuine interactions remains unclear because relevant comparison studies are rare [e.g. 3, 9]. Studies using natural interactions in the home environment should be more reliable in terms of eliciting genuine communicative intent, but they can be extremely time intensive, and the recording set-up must be carefully arranged. Subsequent evaluation of the acoustic analyses must take particular account of this latter factor.

Since these considerations affect choice of experimental set up, a comparative investigation of both natural and laboratory based IDS research is clearly needed. We recently attempted to address this problem by comparing natural speech samples of an existing data set of IDS, ADS and FDS [6], with those produced by students in imaginary interactions [10]. In that study we found that we could reproduce the rated high emotional affective aspects (determined by ratings of low-pass filtered speech) of the different natural speech types in the imaginary laboratory condition (IDS >ADS > FDS). However, we were notably unsuccessful in reproducing hyperarticulation (Formant 1/2 expansion) and the increased pitch normally found in IDS (mean IDS pitch was only higher than mean ADS pitch).

One possible reason for our reproduction of emotional affect may have been the presence of another prosodic characteristic of IDS, namely exaggerated pitch contours. Pitch contours in IDS have routinely been analysed qualitatively using human raters [3], and with a variety of quantitative methods [4, 7, 11]. We have already investigated pitch contour shape in our natural speech samples qualitatively and quantitatively [11], and found that IDS presented more complex and exaggerated pitch contours than both ADS and FDS with no difference between the two adult conditions. This finding was consistent with the direction of the emotional affective analyses in that sample, and here we investigate whether we can replicate the occurrence of exaggerated pitch contours in IDS in speech samples based on imaginary scenarios.

The aim of the present study is therefore to further investigate how comparable imagined and natural speech are by focussing on pitch contours in our existing natural [11] and imagined data sets. If imagined and natural prosodic modifications of IDS are truly comparable we would expect to find broadly similar results in both data sets. To test this we characterise both sets of pitch contours using qualitative human raters, and a previously evaluated [11] quantitative geometric shape analysis technique called Eigenshape analysis (EA). For a rationale of and full introduction to this technique, we refer the reader to our paper in the proceedings of the 2006 Prosody Conference [11].

2. Method

The natural speech samples consisted of 10 southern English mothers (mean age 30.7 years) interacting with their infants, a foreign (Chinese) and southern English confederate (both adult females). For each of the interactions the mothers were provided with three toys to maintain consistency of conversation content, otherwise the interactions were natural and spontaneous. Further information about the mothers, confederates and procedures can be found in Uther et al. [6] and Knoll et al. [11].

2.1. Speech samples

Speech samples consisted of recordings of ten female students (mean age 22.9 years, *sd* 8.84) who imagined talking to an infant, a foreign and a British adult. To keep these imaginary situations consistent with the natural speech samples, the speakers were instructed to imagine that they were talking to a close family member (e.g. niece, nephew or even own child) in the infant condition, but we did not provide speakers with an example or idea of how IDS should sound. For both the British and foreign adult interactions, speakers were instructed to imagine talking to a female stranger in her early twenties. Additionally, in the foreign interactions they were instructed to imagine that the person had been living in the UK for less than two months, and that they might encounter some communication problems. In order to avoid sources of variation due to different utterances (a problem in non-standardised speech samples), we provided the speakers with the same toys as the mothers in the original study to ensure that they used the same target words ‘sheep’, ‘shark’ and ‘shoe’ in each condition. Pitch contours were also extracted from these, because they had been found to be prosodically highlighted in earlier studies [4, 6], and we assumed they would therefore represent excellent comparison stimuli. Apart from the toy stimuli, the speakers were encouraged to construct and invent their own scenarios, which ideally should have resulted in free speech.

2.2. Extraction of pitch contours and image files

A total of 178 words were used for the extraction of the pitch contours (IDS = 58, ADS = 60, FDS = 60). Pitch contours were extracted in Praat 4.1.19, the pitch range for the extraction was set at a floor of 100 Hz and a ceiling of 600 Hz. In order to obtain more homogeneous contours, the ‘smooth’ function was used (set at bandwidth 10 Hz). Finally, the pitch contours were drawn in Praat 4.1.19 (set at a pitch range of 100-600 Hz, duration of 0.7 seconds). Pitch contours were then standardised for duration without distorting the shape, and the lines were then converted to a standard TIFF format at 72 dpi (width 142 pixels) in Corel PhotoPaint 11.0. The reason for this procedure was to ensure a standardised format that could be used in the both qualitative and Eigenshape approaches.

2.3. Procedure

2.3.1. Qualitative analysis

Five raters (three females, two males; mean age 31.2 years, *sd* 5.97) were used to rate and categorise the 178 pitch contour images using a rating scheme with five shapes and one undecided option. The shapes consisted of 1) bell; 2) complex; 3) falling; 4) rising, and 5) level shapes (see additional material), chosen on the basis of previous studies [3]. Raters were informed that the shapes on the rating scheme were ideal representations and that the presented images would probably not exactly resemble these ‘ideal shapes’. Raters were given five trial images to familiarise themselves with the images and rating scheme. Once this trial was completed, the raters were presented with the 178 images in counterbalanced order to avoid order effects. The order was reversed for the last two raters to avoid anomalous results for the final images. In order to determine reliability, intraclass correlation was carried out for the five raters (reliability coefficient $\alpha = 0.881$), and was found to be good. To determine the intra-rater reliability, rater number two repeated the procedure three weeks later, and this was also found to be good (reliability coefficient $\alpha = 0.921$).

2.3.2. Eigenshape analysis

Eigenshape Analysis (EA) uses a principal component analysis-like approach to characterise the actual shape variation of pitch contours through transformation of the contours’ x-y coordinates (detailed discussion of the eigenshape technique would be outside the scope of this paper, but can be found elsewhere [12]). TpsDig 2.0 [13] was used to collect the contour coordinate nodes from the pitch contour TIFF images. The number of coordinate points required to accurately characterise the curve depends on the complexity of its shape, must remain constant in all images and begins at a common landmark point. To keep our study consistent with Knoll et al. [11], we used 37 coordinate points to characterise all curves. The coordinate pairs were then transformed from their Cartesian (x-y) form to a Φ shape function using MacLeod’s ‘X-Y to Phi.exe’ program to provide a series of net angular deviations from the starting point that represents a dimensionless map of the contour shape [12]. Based on the recommended accuracy level of 95% [12], nine nodes were interpolated from the original 37 node curves. This is in contrast to Knoll et al. [11] where 18 nodes were interpolated, which indicates that overall those shapes had been more complex than in the present data set. The Φ-transformed coordinate data output of ‘X-Y to Phi.exe’ was then analysed using ‘Eigenshape.exe’. The output of this program includes determination of each eigenshape, the mean eigenshape, eigenvalues and the individual eigenshape scores on each eigenshape axis (used for further statistical analysis). X-y coordinates that can be plotted for graphic visualisation of each eigenshape and the mean eigenshape were provided by ‘Phi to X-Y.exe’. Eigenscores for each eigenshape were then used for discriminant cannonical variates analysis (CVA). The above programs are freely available at http://www.life.bio.sunysb.edu/morph/.

3. Results

3.1. Qualitative analysis

Chi square analysis indicated that rated pitch contour category and type of speech recipient group variables were associated, ($\chi^2 = 67.761$, df = 8, $p < .001$). Cramer’s V produced a value of .197, which indicates a low relationship between the two variables. Goodman and Kruskal’s Lambda was also calculated for type of speech recipient group ($\lambda = .038$) and rated pitch contour category ($\lambda = .019$), which showed that both variables were not very predictive of each other. This result is also evident in the distribution of ratings for each group (Table 1). A high proportion of IDS contours were characterised as level contours (over 30%) and complex contours (23.8%), whereas bell, rising and falling contours achieved a similar distribution (approx. 15%; see Table 1). This result is in contrast to Knoll et al. [11], where over 60% of the IDS pitch contours were characterised as bell contours, followed by 16.3% level contours (Table 1). Similar to Knoll et al. [11], the majority of ADS contours were characterised as level contours (approx. 40%), although the frequency of level contour ratings was much higher in natural ADS (approx. 80%). Interestingly, the characterisation of bell contours in imaginary ADS is similar to natural ADS and relatively low. In FDS, we found an even distribution of rising, falling and level contours in the imaginary interactions in contrast to the natural interactions. Level contour characterisation was much lower in the imaginary FDS interactions than in the natural FDS interactions, whereas the opposite was true for rising and falling contours. In both ADS and FDS, we found higher occurrences of complex contours, compared to no occurrence of complex contours in the natural speech.

Table 1: *Distribution of ratings in percent for each of the five contours across IDS, FDS and ADS. Values in brackets refer to qualitative analysis of the natural speech samples [11].*

Speech groups	Contour shape				
	Bell	**Complex**	**Rising**	**Falling**	**Level**
IDS	13.8	23.8	15.5	15.9	31.0
	(63.7)	(12)	(4.7)	(3.3)	(16.3)
ADS	2.0	12.0	26.1	20.1	39.8
	(1.1)	(0)	(2.3)	(15.5)	(81.1)
FDS	8.7	16.4	28.4	23.7	22.7
	(1.5)	(0)	(6.3)	(13.3)	(78.9)

3.2. Eigenshape analysis

Almost 54% of the shape variation in all three conditions was attributable to the first eigenshape axis (this information is provided by the programme and indicates the shape variation between groups). This is in contrast to Knoll et al. [11], where almost 70% was attributable to the first eigenshape axis. Separate analysis of each of the three groups demonstrated that the variation on the remaining eigenshape axes was due to all groups. This indicates that the three speech groups are similar in that they possess very little shape variation. Knoll et al. [11] had found that most of the variation of the remaining eigenshape axes was due to IDS (exaggerated contours), whereas ADS and FDS pitch contour variation (level contours) was more similar to each other. EA also provides coordinates for the graphical representation of mean shapes (see Figure 1 for comparison of Knoll et al. [11] and the present data set). The mean shapes of the three groups are relatively similar to each other and comprise an inverted bell curve (Figure 1B), and complement the results of the qualitative analysis. This is probably due to the high occurrence of level, rising and falling contours in each of the three groups, which would result in this shape. Conversely, in Knoll et al. [11], the IDS mean eigenshape was characterised by a more exaggerated curve than either ADS or FDS (see Figure 1A), whereas the mean shape of ADS and FDS were almost identical level contours.

To determine the distinctiveness of the three speech groups, a CVA (simultaneous entry) was performed with speech recipient groups as the independent variable, and the scores on the nine axes derived from EA as the dependent variable. Two discriminant functions were calculated, but none of these was statistically significant. Overall the discriminant functions predicted a successful outcome for 44.4% of all cases, compared to 70.7% in Knoll et al. [11]. Classification for each group was much lower than for the natural speech samples (Table 2). Interestingly, there was an even distribution between the three groups with regards to which group was mistaken for which. For instance over 30% of IDS shapes were mistaken for ADS shapes and 25.86% were mistaken for FDS shapes. The distribution in the adult conditions is similar to this, whereas in the natural speech samples most of the adult conditions were mistaken for each other, and only a small portion of the IDS shapes had been mistaken for either ADS or FDS.

Table 2: *Predictions of discriminant functions for each of the three speech groups. Values in brackets refer to qualitative analysis of the natural speech samples [11].*

Speech groups	Predicted speech groups		
	IDS	**ADS**	**FDS**
IDS	43.10	31.03	25.86
	(75.00)	(15.00)	(10.00)
ADS	30.00	43.33	26.67
	(1.89)	(77.36)	(20.75)
FDS	25.00	28.33	46.67
	(5.56)	(35.19)	(59.26)

Figure 2 shows plotting of each of the pitch contours for IDS, ADS and FDS on functions one and two. Compared to the natural speech samples (Figure 2A), where IDS was very distinct from both ADS and FDS, the current data set (Figure 2B) does not show any distinction between the three speech groups.

4. Discussion

Our results show that, in this instance, the exaggerated pitch contours (compared to ADS and FDS) found in natural IDS were not reproducable in an imaginary laboratory context. As such, our results are consistent with our findings with regards to vowel hyperarticulation in the same sample [10], but do not follow our findings with regards to rated emotional affect[10]. Interestingly, rather than imaginary IDS presenting similar level contours to natural ADS and FDS, the speech groups seemed to have more intra-group variation, and a wider distribution across the shapes than those of the natural sample. The greatest increase in number of shapes seems to have occurred in the rising contour category, which is associated with questioning. It could be that the speakers uttered the target words mainly in a questionning manner.

Our findings differ from those of previous studies investigating *pitch range* in imaginary IDS, FDS and ADS [7, 8], where IDS had been found to contain a greater pitch range than the adult conditions. The reason for this might be due to our experimental set-up.

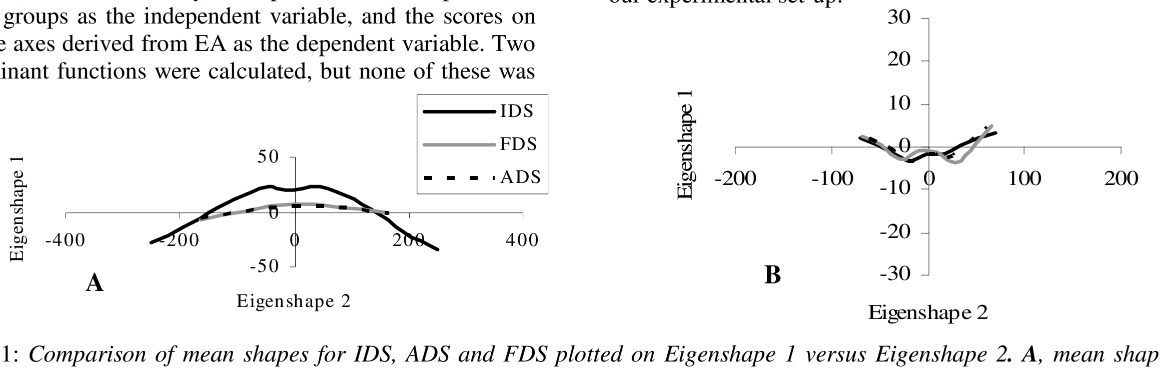

Figure 1: *Comparison of mean shapes for IDS, ADS and FDS plotted on Eigenshape 1 versus Eigenshape 2.* ***A****, mean shapes for natural speech samples [11];* ***B****, mean shapes for present data set.*

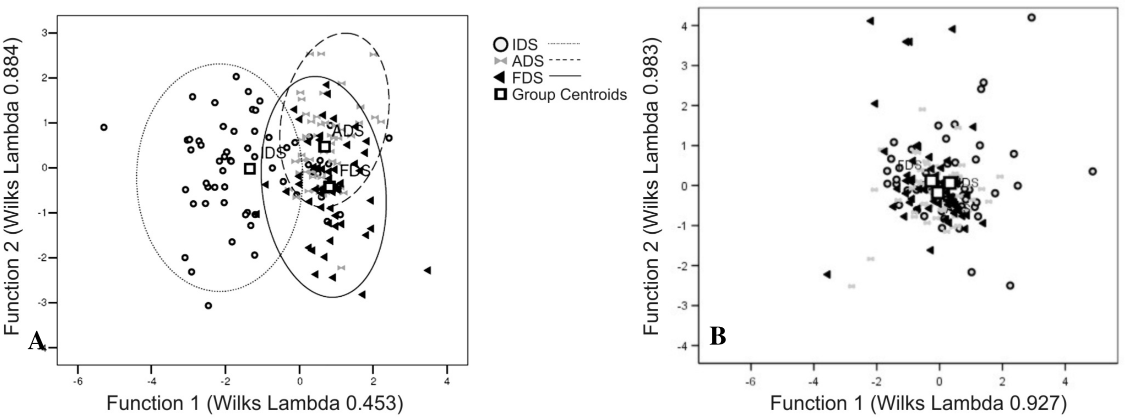

Figure 2: *Combined plot of CVA functions for IDS, ADS and FDS in A, natural speech samples [10] and B, the present data set.*

In contrast to previous studies which used standardised utterances [e.g. 4, 8, 9], we asked our participants to produce free speech without a script and using their own scenarios, although the same stimuli were used to elicit the target words. It is possible that the opportunity to produce spontaneous speech and the need to think about a scenario might have distracted the participants from concentrating on modifying their prosody. We have already collected standardised sentences that remove the need for improvisation from the same participants, and we are currently investigating whether these sentences display the same characteristics as our original natural speech samples. Another way of determining whether the additional task of inventing role play scenarios interfered with the participants' ability to successfully modify their voice is to investigate speakers who are already used to taking part in imaginary role-play activities. To this end, we are currently in the process of investigating the ability of trained actresses to complete the same imaginary tasks.

As previously mentioned, experience with the target listener group may also have a bearing on participants' ability to convincingly mimic a particular speech type. To investigate this we had also asked participants to rate their exposure to each of the speech groups on a five-point Likert scale. These data showed that exposure to infants (mean 2.9) and foreigners (mean 2.6) was relatively low. It is possible that these students' experience was insufficient to effectively mimic these particular groups.

It remains possible that imaginary scenarios are not ideal for eliciting genuine pitch contour modifications. Previous research [e.g. 13] has indicated that the prosodic modifications in IDS are reflexive, instinctive and perhaps even unconscious. If so, experience with infants may not strictly be necessary so long as a real infant is present to interact with. Similarly, a foreign confederate with genuine comprehension problems may be required to elicit clarity-enhancing speech modifications. As such, a real speech partner will provide genuine feedback as part of normal natural speech.

Lastly, because our investigation arose from previous research [e.g. 3, 11] it concentrated on graphical representation of the pitch contours, which is directly related to speech production. Future research should investigate the effect of these pitch contour differences on speech perception. For instance it might be informative to examine whether human listeners may be able to judge which speech condition each of these contours belong to.

5. Conclusions

Our results suggest that natural speech pitch contours (prosody) may not be replicable in imagined speech. Consequently we suggest that imaginary scenarios should be used with caution in studies investigating pitch contour characteristics.

6. References

1. Kitamura, C.; Burnham, D., 2003. Pitch and communicative intent in mother's speech: adjustments for age and sex in the first year. *Infancy* 4, 85-110.
2. Kuhl, P. K., 2004. Early language acquisition: cracking the speech code. *Nature* 5, 831-842.
3. Fernald, A.; Simon, T., 1984. Expanded intonation contours in mother's speech to newborns. *Developmental Psychology* 20, 104-113.
4. Trainor, L.; Austin, C.; Desjardins, R., 2000. Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science* 11(3), 188-195.
5. Burnham, D.; Kitamura, C.; Vollmer-Conna, U., 2002. What's new pussycat? On talking to babies and animals. *Science* 296, 1095.
6. Uther, M.; Knoll, M.; Burnham, D., 2007. Do you speak E-n-g-l-i-s-h? A comparison of foreigner- and infant-directed speech. *Speech Communication* 49, 2-7.
7. Biersack, S.; Kempe, V.; Knapton, L., 2005. Fine-tuning speech registers: a comparison of the prosodic features of child-directed and foreigner-directed speech. *9th European Conference on Speech Communication and Technology.* Lisbon, 2401-2402.
8. Papousek, M.; Hwang, S-F., 1991. Tone and intonation in Mandarin babytalk to presyllabic infants: comparison with registers of adult conversation and foreign language instruction. *Applied Psycholinguistics* 12, 481-504.
9. Schaeffler, F.; Kempe, V.; Biersack, S., 2006. Comparing vocal parameters in spontaneous and posed child-directed speech. *3rd Conf. on Speech Prosody.* Dresden, 688-691.
10. Knoll, M. A.; Scharrer, L., 2007. Acoustic and affective comparisons of natural and imaginary infant-, foreigner- and adult-directed speech. *9th Interspeech Conference.* Antwerp, 1414-1417.
11. Knoll, M.; Uther, M.; MacLeod, N.; O'Neill, M.; Walsh, S., 2006. Emotional, linguistic or cute? The function of pitch contours in infant- and foreigner-directed speech. *3rd Conf. on Speech Prosody*. Dresden, 165-168.
12. MacLeod, N., 1999. Generalizing and extending the eigenshape method of shape space visualisation and analysis. *Paleobiology* 25(1), 107-138.
13. Papousek, M.; Papousek, H.; Bornstein, M. H., 1985. The naturalistic vocal environment of young infants: on the significance of homogeneity and variability in parental speech. In *Social perception in infants,* T. M. Field; N. A. Fox (Eds.). Norwood: Ablex, pp.269-297.