Vowel data of early speech development in several languages

Louis C.W. Pols[1], Elena Lyakso[2], Jeannette M. van der Stelt[1], Ton G. Wempe[1] & Krisztina Zajdó[3]

[1)] Institute of Phonetic Sciences / ACLC, University of Amsterdam, The Netherlands
[2)] Saint-Petersburg State University, St.-Petersburg, Russia
[3)] University of Wyoming, Division of Communication Disorders, Laramie, WY, USA

Louis.Pols@uva.nl

Abstract

It is notoriously difficult to perform reliable spectro-temporal analyses on speech of young children (up to two years of age), partly because of the high pitch of their voices. Formants are very poorly defined and thus we used a pitch-synchronous bandfilter analysis, followed by a principal components analysis to represent the acoustic characteristics of children's vowel(-like) productions. Since young children's vowel realizations are very hard to label consistently, it is fortunate that this analysis method works just as well for unlabeled data. This way we can still study the size and form of the acoustic vowel space and its changes over the first two years of life. This is indispensable information to develop a universal vowel acquisition theory and to establish effective treatment methods to remediate disordered vowel productions. So far we have analyzed vowel data from 5 to 8 boys at two years of age in each of the following three languages: Dutch, Hungarian and Russian. Many practical difficulties were encountered (related among other things with the ways of collecting and actually recording the vocalizations), that made comparisons within and between languages not always easy, but nevertheless some language specific properties appear to be detectable.

1. Introduction

We limit ourselves in this study to the segmental quality of speech, thus neglecting any suprasegmental, prosodic or linguistic aspects. Vowels from the three languages studied here, Dutch, Hungarian and Russian, substantially differ from each other phonemically. Differences include the number and properties of short and long vowels (whether or not diph-thongized), the existence of true diphthongs, palatalization, etc. New-borns potentially have the capability to learn to produce any vowel set present in their environmental language(s), if properly stimulated. Still, certain sounds are acquired later than others, probably because of articulatory difficulties.

One would like to compare the natural development of a young child with certain norms and landmarks, in order to be able to determine whether a developmental delay occurs and whether more attention or specific treatments are required. Once the baby starts to produce interpretable word-like utterances, it is easier for adults to judge certain qualities of the child's speech. However, also during the first year of life it would be challenging to do so. One possibility to transcribe speech-like utterances, such as babbling, would be in terms of respiration, phonation and articulation [2]. In principle, it should also be possible to physically analyze these early vocalizations and describe them in terms of dynamically varying pitch and formant characteristics and label them as phoneme-like segments. Actually this is a very difficult task with a very low consistency. Many good reasons can be given for this. The pitch of these child voices can be extremely high, up to 400 Hz and even higher, with much variation. Thus, the line spectrum is poorly defined. This implies that also the formant frequencies are poorly defined. It is very well possible that young baby's produce one- or two-formant sounds only, given that their oral cavity is largely filled with the tongue. It is our impression that if nevertheless formant values are measured, those are strongly influenced by the expectations of the researcher. For these and other reasons we prefer to use a more perception-oriented approach based on a bandfilter analysis followed by some from of data reduction. For adult speech, Pols et al. [3, 4] have shown that an objective and automatic formant-like vowel representation can be derived by using some 18 one-third octave filters, similar to the critical band filters of the inner ear. One spectral measurement then results in set of 18 bandfilter values expressed in dB. These 18 numbers can also be represented as a point in an 18-dimensional space. Subsequent measurements in time within one utterance, or additional measurements of other vowels, then result in multiple points in that space. Since the variability of all possible vowel spectra is rather limited, this 18-dimensional space can substantially be reduced in dimensionality without hardly any loss of information. One way to achieve this data reduction is by applying a principal components analysis that is based upon maximizing the amount of variance explained. One of the beauties of this approach is that no labeled data are required. So, it is also applicable for unlabeled baby vocalizations. It appears that such a 2- or 3-dimensional vowel representation shows compelling similarity with a formant representation, whereas also the perceptual similarities between vowels are nicely represented. Of course, also this approach is far from perfect, but it is objective and can be performed automatically. Below we will present our present procedure as well as its limitations, but first we will quickly introduce the child recordings that we had available for our analyses.

2. Child speech recordings

In this study three pre-existing speech data sets for Dutch, Hungarian and Russian children were analyzed. Ideally, it is desirable to work with fully comparable material, but that was not the case here. So, for the time being we will have to accept differences in recording quality, in elicitation procedures (free communication in home situation *vs.* controlled interaction in the lab), and otherwise. In this presentation we will limit ourselves to the speech of two-

year-old children, although material at younger and/or older age are available as well.

The Dutch and the Russian speech material (5 boys each) was recorded in a home setting while the children conversed with their mothers. The Hungarian speech data (8 boys) were collected in the lab during naturalistic interactions between a child and its caregiver. 28 Puppets with pre-assigned CVCV-structured names (such as /pi:pi:/ or /tu:tu:/) were modeled by the caregiver to elicit imitations from the child.

50 Individual utterances per Dutch child were selected randomly, resulting in 250 utterances. For Hungarian, 763 imitated puppet name utterances were available from which 229 were selected. Also for the Russian data set 50 utterances per child were collected. For more details see [6, 7, 9].

All subsequent analyses were performed by *praat* [1] and were automatically controlled by scripts. Potentially 10 spectral measurements were performed spread over each utterance. However, certain physical criteria had to be met. The selected vocalic segments should neither be clipped nor be too low in level (between 0,5 and 10 dB below the absolute peak level within the utterance), they should be voiced but the pitch should not exceed 425 Hz. For all those segments per utterance for which all criteria were met, the analysis algorithm was applied. For more details see [7].

3. Spectral analysis method

The pitch-synchronous bandfilter analysis method (for more details see [8]) includes the following steps. For each segment that met all selection criteria the nearest F0 period in the middle of the segment was selected. This period then was recycled up to a duration of 50 ms and multiplied with a Kaiser-window to reduce the 20-Hz ripple. Additionally, pre-emphasis was applied. Next, a swept Gaussian bandpass filter analysis was performed (step=175 Hz resulting in 40 filters between 0 and 7000 Hz; effective fixed bandwidth = 1.1 x 425 Hz) and a level normalization was applied. Due to the selection criteria for pitch and intensity, on average about 5 spectral measurements per utterance were performed, rather than the maximum number of 10. Each spectrum consists of 40 linear intensity values, the data are thus 40-dimensional and can be reduced by applying a principal components analysis.

4. Principal components analysis

In a principal components analysis (PCA) subsequent perpendicular new dimensions are selected (being linear combinations of the original ones, specified by the direction cosines of the eigenvectors), that explain as much of the (remaining) variance as possible. Thus, the (variance in the) set of unlabeled points itself defines the new lower-dimensional representation. This property of the method has certain consequences when comparing data over languages. To build up a 2- or 3-dimensional reference space for the vowel data of two-year old children, we have decided to start from comparable sets of data per language in terms of size. An equal number of 980 level-normalized vowel spectra per language were identified and projected together in one joint space. A PCA *per language* would produce a reference representation optimized per language, but would make comparison over languages much more difficult. Most of the time more than 980 spectra per language were available, from which then a random selection from each child was taken. Since the recording conditions were not identical, one could decide to make a correction for that as well, for instance by subtracting the overall spectrum per child from all its spectra, called *centering* in [4]. However, this procedure indirectly also normalizes away a possible effect of the distribution of the vowels in the set, so we decided not to do that. The combined Dutch-Hungarian-Russian reference space is defined by a set of eigenvectors of which the first two (ev1 and ev2) are presented in Fig. 1. The corresponding percentages of variance explained are given in Table 1. The pc1-pc2 plane of the combined Dutch-Hungarian-Russian data set (3 x 980 points) is presented in Fig. 2 on the last page. In subsequent figures language-specific data are projected in this reference plane.

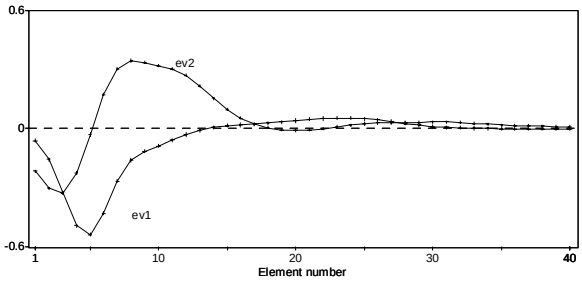

Fig. 1: First two eigenvectors of the joined Dutch-Hungarian-Russian vowel space.

Table 1: Percentage of variance explained by the first four eigenvectors, both individually as well as cumulatively.

eigenvector	% variance	% cum. variance
1	27.95	27.95
2	25.62	53.57
3	13.37	66.94
4	9.0	75.94

5. Language-specific data

Results in the previous section were based on unlabeled data. However, at two years of age it is already possible to consistently label certain segments in the selected utterances of good quality as native vowels. For the material in the three different languages 15 (when available) correctly produced /i:/, /u:/, and /a:/ corner vowels (according to the judgment of an adult native listener) are presented in Figs. 3-5 as a reference.

It is tempting to interpret the pc1-pc2 plane as a (perhaps somewhat rotated) F1-F2 formant plane. In [4] it was shown that for average adult vowel data and for the bandfilter analysis used in that paper, such an excellent fit could be achieved.

Since we are talking here about an evolving vowel system that may not be fully developed yet at two years of age, it is interesting to examine the overall size and the local distributions of the vowel spaces for these two-year olds in the three different languages. And actually even more so to study the development over age per language from the first recordings available (for the Dutch children this is 4-6

months). However, this approach requires some form of cluster analysis or vector quantization on unlabeled data that, unfortunately, we are not yet aware of. To characterize the overall size of the vowel space, we use one-standard deviation ellipses. It is clear that an ellipse might not be most appropriate for describing the distribution of points in a triangularly- or quadrangularly-shaped field.

In Figs. 3-5 (see next page) all 980 vowel data points per language are displayed with the one-standard deviation ellipse around the mean. Three smaller ellipses represent the distributions of a subset of (most of the time) 15 corner vowels /i/, /u/ and /a/ that were perceived as correctly produced. In Table 2 some additional details of the data per language are presented.

Table 2: Some details of data per language

	Dutch	Hungarian	Russian
number of boys	5	8	5
nr. of utterances	5 x 50	763 → 229	5 x 50
total nr. of segments	1392	988	980
nr. of segments used	980	980	980
identified /a/, /i/, /u/	76/54/68	98/110/127	6/29/19
displayed /a/, /i/, /u/	15/15/15	15/15/15	6/15/15

It will be clear from Figs. 3-5 that the differences between the three languages are substantial. The 980 Dutch data points in Fig. 3 cover a large part of the pc1-pc2 space. Dutch labeled corner vowels vary more and are also more peripherally located than in the other two data sets. The Hungarian and Russian results are more alike in terms of the positioning of the high corner vowels which are closer to each other than in Dutch. Lip rounding may account for the varying degrees of closeness of the high corner vowels. Although interpreting these differences in terms of language-specific phenomena might be possible, careful consideration of potential artifacts is essential.

The labeled corner vowels have been identified by a native speaker in each language. It is very well possible that one expert listener was more strict than the other and less willing to include a vowel segment if it could not be identified as such in isolation but only within the context of the whole utterance. From the Dutch expert it is known that she accepted any vowel segment as long as it was identifiable within the context of the whole utterance. The Hungarian listener limited herself to long vowels only.

Another technical complication is the highly variable quality of the recordings within and between languages. A home recording made with a fixed microphone position while the child moves around, leads unavoidably to far-from-perfect sound quality. Precautions were taken to exclude recordings at very low level, but most of our data are far from ideal in this respect.

6. Discussion and future work

Since comparable data sets for child speech in other languages than Dutch, Hungarian and Russian are available from other researchers, we hope to be able to extend our data with several more languages. With more languages used to build up a reference space at two years of age, it is expected that data from additional new languages and from children at other ages, can simply be projected in that space, without having to recalculate a new reference space for this larger set of languages. Another reference space can be created by using *adult* vowel data. The topic of speaker normalization then becomes apparent, especially when comparing male and female data [5].

Although we have presented here data from 2-year old children, it might actually be more informative to analyze speech of younger children, for instance to be able to say more about the supposed strong influence of the mother tongue already at four months of age. Vocalizations of such very young children are notoriously difficult to analyze spectrally, but we have good hope that our approach will be successful here as well.

This study focused on presenting spectral information from individual segments taken from utterances. However, subsequent segments *within* an utterance can also be presented as a series of points in that reference space, which would allow to study dynamic spectral behavior such as diphthongization.

As already indicated in the previous section, we are in need of procedures to define in an objective way the overall size as well as the local distributions of unlabeled sets of points. Arguably, vector quantization techniques as used in speech coding or in automatic speech recognition could be potentially helpful for this purpose.

7. References

[1] Boersma, P. & Weenink, D. (1996), "PRAAT, a system for doing phonetics by computer, version 3.4", *Institute of Phonetic Sciences of the University of Amsterdam, Report* 132, 182 pp., see also www.praat.org.

[2] Koopmans-van Beinum, F.J. & van der Stelt, J. M. (1986), "Early stages in the development of speech movements", In: Lindblom, B & Zetterström, R. (Eds.), *Precursors of early speech*, Stockton Press, New York, 37-50.

[3] Pols, L.C.W., van der Kamp, L.J.Th. & Plomp, R. (1969), "Perceptual and physical space of vowel sounds", *J. Acoust. Soc. Amer.*, 46, 458-467.

[4] Pols, L.C.W., Tromp, H.R.C. & Plomp, R. (1973), "Frequency analysis of Dutch vowels from 50 male speakers", *J. Acoust. Soc. Amer.*, 53(4), 1093-1101.

[5] Pols, L.C.W. & Weenink, D.J.M. (2005), "Vowel recognition and (adaptive) speaker normalization", *Proc. SPECOM 2005*, Vol. 1, 17-24.

[6] Stelt, J.M. van der, Lyakso, E. & Gromova, A. (2005), "Two-year-old Dutch- and Russian-speaking children: Exploring the vowel space", *Proc. 10th Int. Congress for the Study of Child Language*, Berlin, 159.

[7] Stelt, J.M. van der, Zajdó, K. & Wempe, T.G. (2005), "Exploring the acoustic vowel space in two-year-old children: Results for Dutch and Hungarian", *Speech Communication*, 47 (1-2) 143-159.

[8] Wempe, A.G. & Boersma, P. (2003), "The interactive design of an F0-related spectral analyser", *Proc. ICPhS*, Barcelona, Vol. 1, 343-346.

[9] Zajdó, K., van der Stelt, J.M., Wempe, T.G. & Pols, L.C.W. (2005), "Cross-linguistic comparison of two-year-old children's acoustic vowel spaces: Contrasting Hungarian with Dutch", *Proc. Interspeech 2005*, Lisbon, 1173-1176.

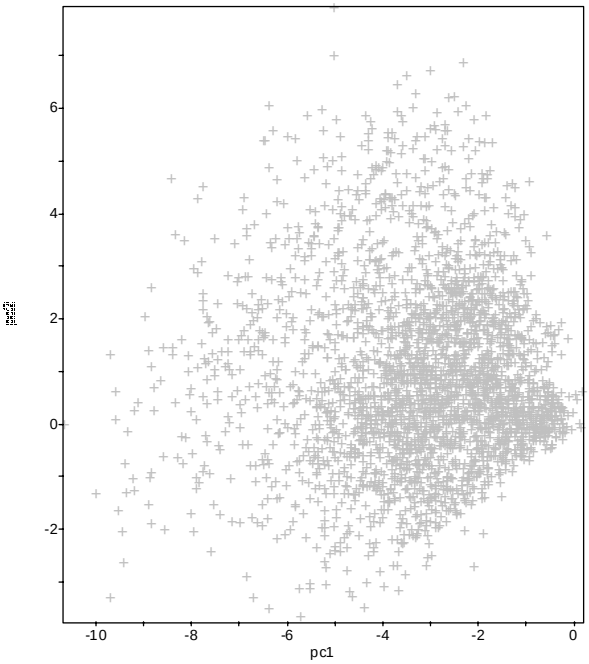

Fig. 2: Reference plane composed of 3 x 980 vowel spectra of Dutch, Hungarian and Russian 2-year old children.

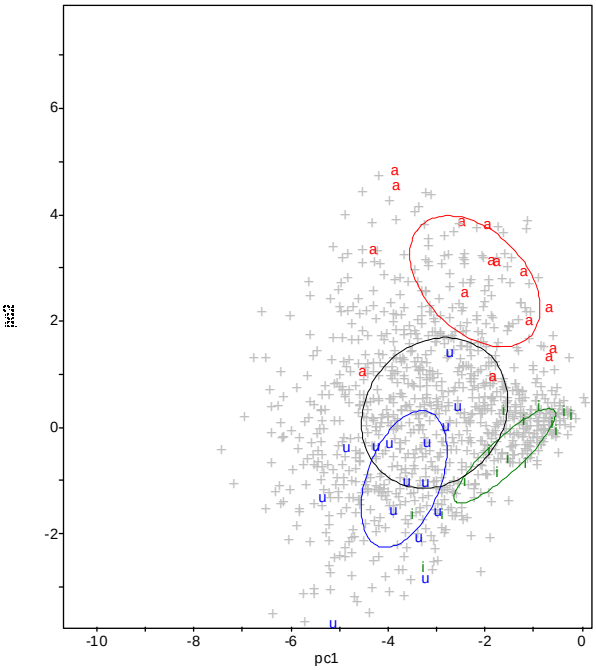

Fig. 4: Hungarian vowel spectra in the joint reference plane. The central ellipse reflects one standard deviation of all 980 Hungarian vowel spectra. The blue, red and green ellipses represent one standard deviation of a subset of labeled long /u/, /a/ and /i/ vowels, respectively.

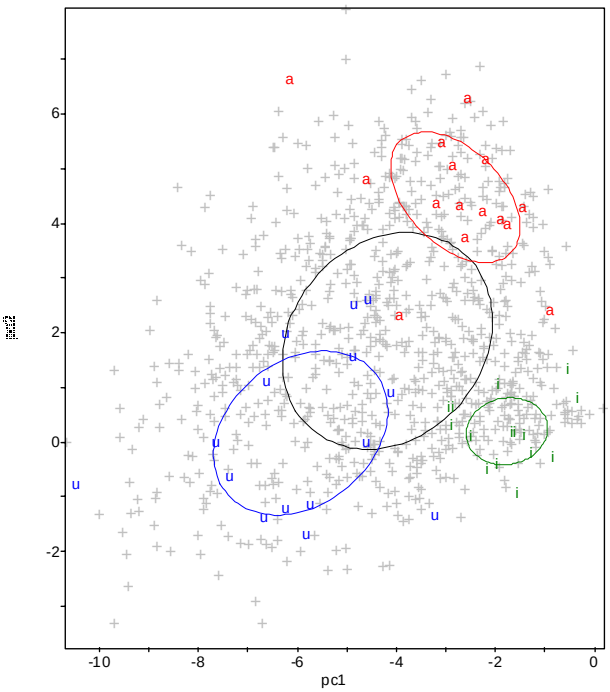

Fig. 3: Dutch vowel spectra in the joint reference plane. The central ellipse reflects one standard deviation of all 980 Dutch vowel spectra. The blue, red and green ellipses represent one standard deviation of a subset of labeled /u/, /a/ and /i/ vowels, respectively.

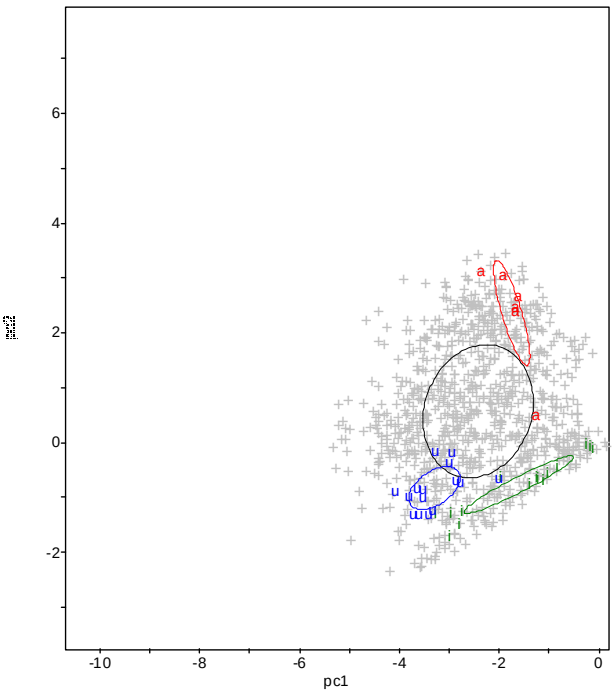

Fig. 5: Russian vowel spectra in the joint reference plane. The central ellipse reflects one standard deviation of all 980 Russian vowel spectra. The blue, red and green ellipses represent one standard deviation of a subset of labeled /u/, /a/ and /i/ vowels, respectively.