# Mothers Reveal More of Their Vocal Identity When Talking to Infants

T. Kathiresan[1,9,11], A. Hervais-Adelman[2,7,8], S. Townsend[3,4,7], L. Dilley[5], R. Shi[6], M. M. Daum[2,7,8,10], V. Dellwo[1,7,11] *

[1]Department of Computational Linguistics, University of Zurich, Switzerland.

[2]Department of Psychology, University of Zurich, Switzerland.

[3]Department of Comparative Linguistics, University of Zurich, Switzerland.

[4]Department of Psychology, University of Warwick, United Kingdom.

[5]Department of Communicative Sciences & Disorders, Michigan State University, USA.

[6]Department of Psychology, University of Quebec in Montreal, Canada.

[7]National Competence Centre of Research *Evolving Language*, Switzerland.

[8]Neuroscience Center Zurich, University of Zurich and ETH Zurich, Switzerland.

[9]Telepathy Labs, Zurich.

[10]Jacobs Center for Productive Youth Development, University of Zurich, Switzerland.

[11]Centre for Forensic Phonetics & Acoustics, University of Zurich, Switzerland.

*Correspondence to: volker.dellwo@uzh.ch

**Summary**

Voice timbre – the unique acoustic information in a voice by which its speaker can be recognized – is particularly critical in mother-infant interaction. Correct identification of vocal timbre is necessary in order for infants to recognize their mothers as familiar both before and after birth [1–6], providing a basis for social bonding between infant and mother. The exact mechanisms underlying infant voice recognition remain ambiguous and have predominantly been studied in terms of cognitive voice recognition abilities of the infant. Here, we show – for the first time – that caregivers actively maximize their chances of being correctly recognized by presenting more details of their vocal timbre through adjustments to their voices known as infant-directed speech (IDS) or *baby talk*, a vocal register which is wide-spread through most of the world's cultures [4,7,8]. Using acoustic modelling (*k*-means clustering of Mel Frequency Cepstral Coefficients) of IDS in comparison with adult-directed speech (ADS), we found in two cohorts of speakers - US English and Swiss German mothers - that voice timbre clusters of in IDS are significantly larger to comparable clusters in ADS. This effect leads to a more detailed representation of timbre in IDS with subsequent benefits for recognition. Critically, an automatic speaker identification using a Gaussian-mixture model based on Mel Frequency Cepstral Coefficients showed significantly better performance in two experiments when trained with IDS as opposed to ADS. We argue that IDS has evolved as part of an adaptive set of evolutionary strategies that serve to promote indexical signalling by caregivers to their offspring which thereby promote social bonding via voice [9] and acquiring linguistic systems [10].

## RESULTS

A predominant challenge in voice recognition relates to the immense variability in timbres of voices within individuals, arising from a large spectrum of sources, such as varying registers, vocal effort or different emotions [11–19]. This variability leads to reduced recognition performance for a voice learned in one register and presented in another. IDS is a register which shows some of the highest within- and between-speaker acoustic variability (see review in [20]). Recently, Piazza et al. [4] demonstrated consistent within-speaker shifts in timbre from ADS to IDS, which they interpreted as changes in the unique fingerprint within a caregiver's voice. They argued it helps the infant differentiate between ADS and IDS. However, if adults indeed shifted their unique vocal fingerprint from ADS to IDS, this should correspond to a perceived change in voice identity with detrimental consequences for recognition performance in cross-register recognition (for example, when a voice is learned under IDS and recognition takes place under ADS). There is, however, consent that IDS registers are applied to support attention and thus facilitate social bonding between caregiver and infant [21–24]. Social bonding, however, requires a solid recognition of the social interactants, notably by their voices [1–6]. We contend that it is biologically implausible that caregivers make recognition challenging by addressing infants in a register that will not serve to make them more generally recognizable. Additionally, IDS features some of the highest within-register timbre variability of all vocal registers, manifested as tremendous exaggerations of acoustic segmental [10,25–28] and prosodic characteristics [29–34]. While within-person variability – within or between register – has generally been thought to oppose identity recognition [35–37], results from automatic voice recognition [18] and – more recently – from human voice recognition [37–40] revealed that learning identity from voices containing more within-speaker detail, can generalize better to unknown variability in the long-term and thus lead to more robust recognition. It thus seems plausible that the high degree of within-speaker variability observed in IDS might contribute positively to recognition when a system has been trained with it. Here, we tested this hypothesis using acoustic modelling and automated speaker recognition models. To test this, we analysed speech from mothers of two independent cohorts (10 Swiss-German and 27 US-English) reading sentences to their

own infants (IDS) and to an adult experimenter (ADS), in recording facilities located in the Upper Midwest, USA and University of Zurich/Switzerland respectively (see STAR methods). Given that infants exhibit a preference for the mother's over the father's voice[69] and that they show a lack of discrimination between male voices (including the paternal voice[70]), we have – in line with previous studies on IDS – limited our investigation to mothers' voices.

**Acoustic modelling: ADS is a timbre subspace of IDS**

To better understand the acoustic nature of the timbral shift suggested by Piazza et al. [4], we modelled this shift in a multidimensional timbre space. The timbre of a voice that forms its identity consists of a set of multidimensional acoustic variables that change rapidly over time. A typical way of modelling timbre is by using Mel Frequency Cepstral Coefficients (MFCCs, [41]), a series of speech coding techniques arriving at a discrete number of coefficients (typically 13) specifying voice-specific timbre during a specified time unit (typically 25 ms). Clusters of MFCCs in the multi-dimensional space are an efficient way of identifying speaker-specific timbre, henceforth *timbre clusters*. To understand how an ADS timbre space of a speaker compares to the IDS space of the same speaker, we first identified 32 timbre clusters in a 13-dimensional MFCC space using *k*-means clustering. We applied the clustering to ADS and IDS combined for each speaker. 32 clusters were chosen because this was the largest number for which each cluster contained sufficient data samples in both IDS and ADS from the same speaker to measure 95% confidence ellipses in dimension-reduced data [42]. Second, we divided each cluster into ADS and IDS sub-clusters and measured the Euclidian distance between the sub-cluster centroids in the 13-dimensional MFCC space. Fig. 1. (A) illustrates the median distance between IDS and ADS in the timbre space, which was about four times higher than the distance between a random data-split with 5-fold cross-validation at the proportion of the ADS/IDS ratio within a cluster; a linear mixed model showed a significant effect of split type, $F[1,1] = 1425.21$, $p < .001$, but not of language, $F[1,1] = 0.025$, $p = .87$; and no interaction: $F[1,1] = 0.33$, $p = .56$). This

shows that, within a timbre cluster, ADS and IDS are centred around different means, consistent with the proposal of a shift in Piazza et al. [4].

To compare the relative extent of the two subclusters in the timbre space, we obtained the variance of the ADS and IDS subgroups within each cluster by applying Principal Component Analysis (PCA) on the MFCC data. Using the first two principal components, we measured the area for IDS and ADS data by fitting a 95% confidence ellipse. Fig. 1. (B) shows the area of ADS and IDS sub-clusters in a representative example of one single *k*-means cluster of a Swiss speaker; areas of IDS and ADS in all clusters for both languages are summarized in Fig. 1. (C). Both languages show a proportionally smaller area for ADS compared to IDS (pairwise t-test with Bonferroni correction for Swiss: $p = 0.0019$; US: $p < 0.0001$). The difference was significantly larger for US English compared to Swiss German, as revealed by a significant interaction between register and language (stemming from a linear mixed model [register*language] with [speaker] as random factor: $\beta = -0.9641$; ANOVA: $F[1, 2329] = 9.3711$; $p = 0.00223$).

In addition to demonstrating a method by which the timbral shift suggested by Piazza et al.[4] can be rigorously quantified, the above results crucially identify and confirm a systematic increase in the extent of size of the timbre space occupied by IDS register compared to ADS register. This systematic increase of the timbre space in IDS compared to ADS holds across two distinct languages/dialects – Swiss German and US Midwestern English – involving very disparate geographic regions and populations of individuals. Other than a mere shift between ADS and IDS – as argued by Piazza et al.[4] – this means that the potential vocal space a speaker occupies is better represented in IDS compared to ADS registers, and further, that IDS includes the timbre space of ADS significantly more than the other way around. We thus obtained a rigorous acoustic model that explains the variability advantage demonstrated in prior studies in both computer [19] and human [38] voice recognition.

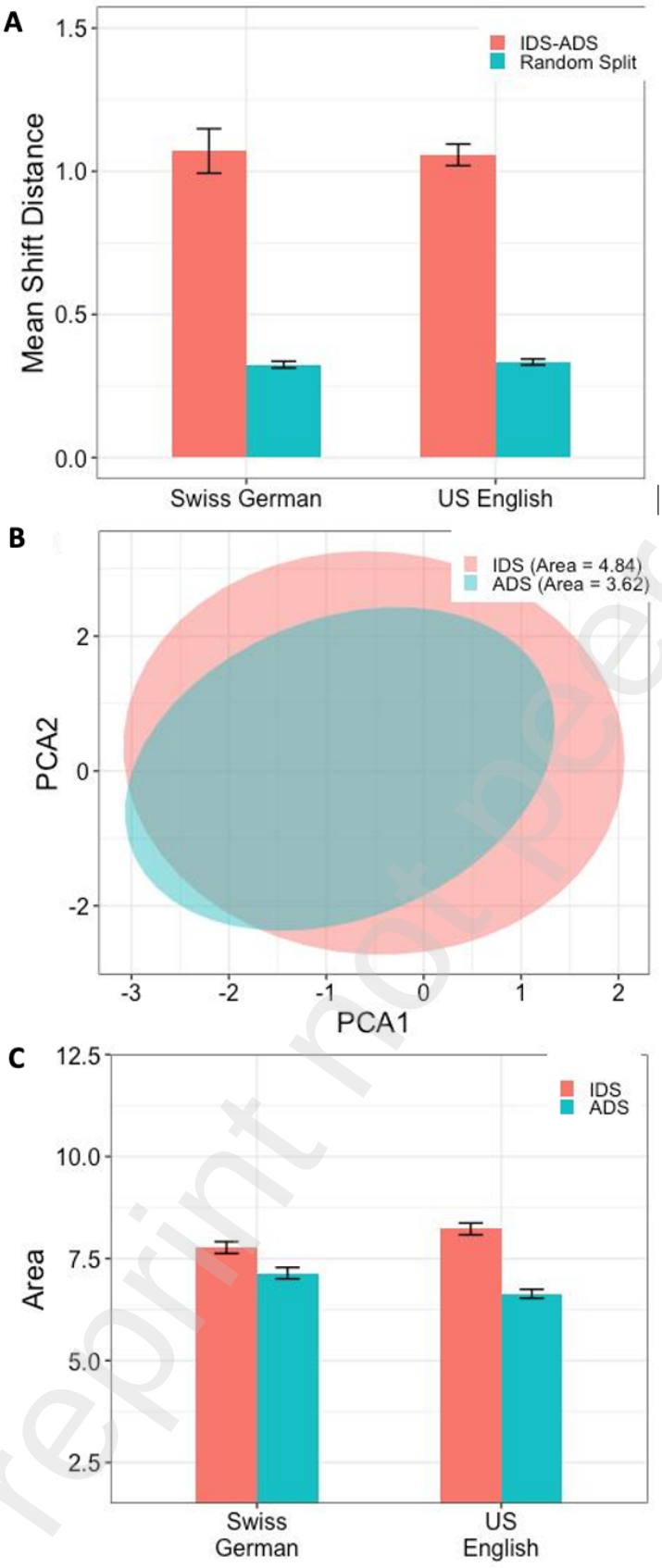



**Figure 1.** **(A)** Within-cluster differences between ADS and IDS centres (red) and differences between random within-cluster splits (turquoise). **(B)** Representative example of IDS (red) and ADS (turquoise) areas in one GMM of one speaker in Swiss German, **(C)** mean *k*-means cluster area (averaged +- 1 stdev over 10 and 27 speakers respectively).

**Expanded timbre space leads to a recognition advantage**

The finding of an expanded timbre space in IDS compared to ADS has direct implications for voice recognition. Because IDS contains more timbre detail compared to ADS, we hypothesised that learning a voice under the larger IDS timbre space should lead to better recognition performance of voices under ADS compared to learning a voice under ADS and recognizing the speaker under IDS (henceforth: IDS-ADS and ADS-IDS respectively). Recognition was tested based on a Gaussian-Mixture Model (GMM) with 32 mixtures constituting the number of mixtures for which almost no empty clusters in either IDS or ADS were obtained, along with 5-fold cross-validation based on 13-dimensional MFCCs (see Fig. 2). A two-way ANOVA [accuracy by language and speaking-register] had a significant interaction between the effects of language and speaking-register [$F\,[1,16] = 80.92$, $p_{(\text{Bonferroni})} < 0.001$], indicating effect magnitudes were somewhat larger in US English compared to Swiss German. Testing the simple effect of training-test sequence (IDS-ADS; ADS-IDS) revealed significantly better IDS-ADS recognition in both languages, as shown by a pairwise *t*-test with Bonferroni correction for Swiss: $p = 0.001$; US: $p < 0.001$). Results were replicated (see STAR methods) for different numbers of mixtures (32 to 512) and including temporal features (first- and second-order derivatives; delta and delta-delta). To make sure that the model performs well under within-register conditions, we also tested ADS-ADS and IDS-IDS. Mean within-register recognition led to ceiling effects for ADS (Swiss: 100%, US: 98%) and close to ceiling performance for IDS (Swiss: 99%, US: 88%). This was to be expected as GMMs perform highly on within-register read speech for our group sizes. We relate the decrease in performance within IDS recognition – particularly for US-English speakers – to the generally higher variability in both training and test data.

Results suggest that learning vocal identity from acoustically more variable IDS may lead to speaker recognition benefits in other, less variable registers (here ADS), but not vice versa. While using less variable registers in training and testing (e.g., read ADS-ADS) leads to optimal results, this scenario is not very applicable, as listeners would only be prepared to identify speakers under this particular low-

variability condition, which is improbable to occur. In real-life situations of kinship recognition, listeners task is to identify speakers under a wide range of within-speaker timbre variability. Thus, by addressing an infant in IDS, the infant may learn critical information about timbre features that serve to identify speakers in ADS and possibly other registers, but the reverse does not hold. It supports the view that the higher amount of attention infants pay to caregivers under IDS [21–24] is accompanied by caregivers revealing more information about their voice timbre variability which in return leads to the acquisition of a more robust representation of that individual's voice. In sum, it suggests that the acoustics of IDS play a crucial role in furnishing the infants with the necessary cues for deriving a highly-generalisable speaker identification system.

The fact that the impact of register on both recognition performance (Fig. 2) and timbre differences (Fig. 1) was higher in US English compared to Swiss German may potentially be explained as differences in the circumstances of IDS/ADS production across the recording sessions in different countries/languages (Swiss and US). In comparison to American mothers, who spoke to adults in the absence of their infants, Swiss German mothers carried their baby while they were producing adult directed speech and read a children's story to the adults. Even though Swiss listeners showed a high sensitivity in differentiating IDS from ADS in the Swiss speakers (mean A' across 31 listeners = 0.79; see STAR methods), Swiss ADS had notable IDS influences. In addition, differences between ADS and IDS registers are continuous rather than categorical, and specific acoustic attributes and prevalence differ according to cultural background, geographic region, and individual identity [43–45]. We therefore posit that the reason for the smaller effect magnitudes in Swiss speakers may also plausibly reflects the use of a generally less marked IDS register by the Swiss German speaker group compared to the US group.

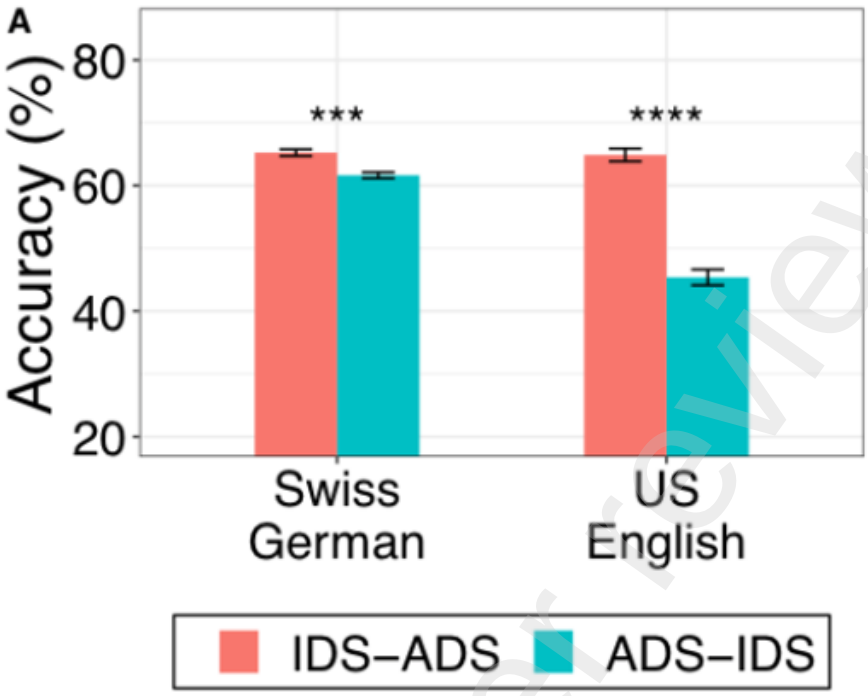



**Figure 2.** Recognition accuracy for GMM models trained on ADS and tested with IDS (ADS-IDS) and the reverse (IDS-ADS) for both US-English (N=27) and Swiss-German (N=10) speakers (average +- 1 stdev over 5 folds of cross-validation).

**Learning a speaker's voice from IDS results in a recognition advantage in other registers**

In a second recognition experiment we used an identical test set to probe recognition performance using the IDS and ADS recognition models from the cross-register condition (previous section) to recognize spontaneous speech elicited in a picture description task and an interview situation. This register was available for the Swiss German speakers only. Both tasks were carried out under IDS and ADS conditions (see STAR Methods). This speech was not subject to the typical constraints of read speech and contained a large variety of different lexical items and grammatical patterns. Fig. 3 (A) shows that the entire timbre space (13-dimensional MFCCs) in a dimension reduced PCA is more extensive than, and mostly occupies a different space to, the read speech of the previous experiment. Recognition performance showed a significant interaction between training set (IDS, ADS) and

spontaneous speech type (free, picture-naming) on accuracy [two-way ANOVA accuracy*test set: F [1,16] = 25.03, $p < 0.0001$]; (Fig. 3. B). Simple effects (Bonferroni corrected p-values) of training register for each of free and picture-naming speech revealed significantly higher recognition performances for IDS compared to ADS-trained models (free-speech: $p < 0.043$; picture-naming: $p < 0.0001$). These results provide strong evidence for the view that IDS compared to ADS training generalizes better to vocal registers that are not part of either of the training registers. Hence, the timbre variability obtained in IDS is likely to be of advantage to recognize voices under a wide variety of within-speaker timbre variants. We thus conclude that training a listener with IDS instead of ADS will prepare them to recognize the speaker under possibly all vocal settings and registers of that speaker.

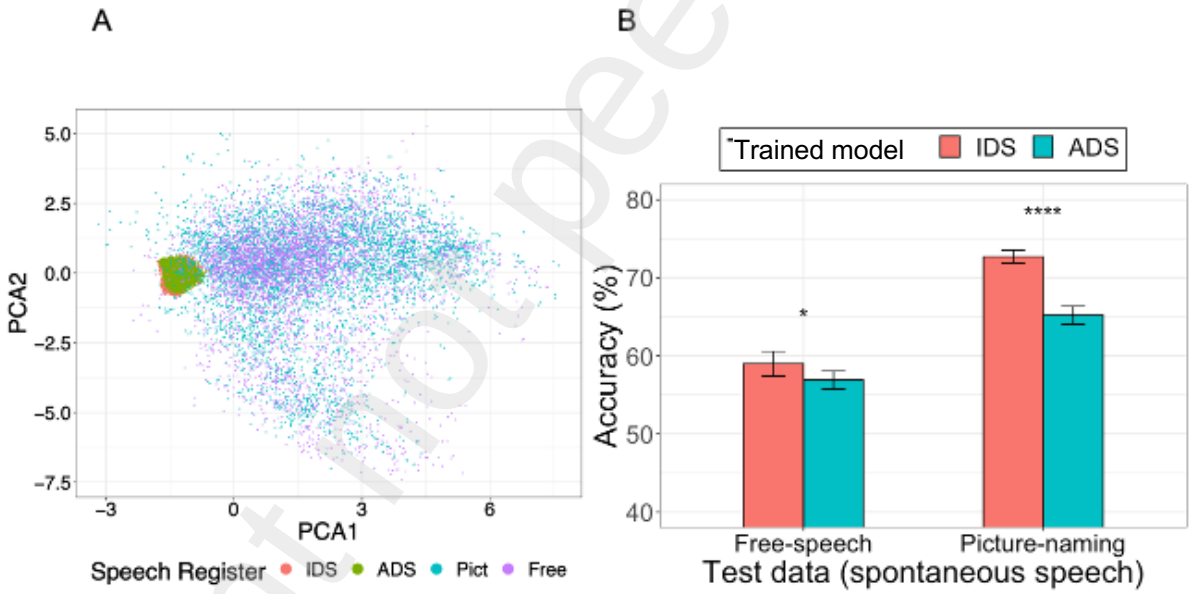



**Figure 3. (A)** Example of read (ADS, IDS) and spontaneous speech (Picture naming, and free-speech) of one Swiss-German mother in a PCA space. Both spontaneous registers occupy a different timbre space compared to the read registers **(B)** The trained IDS and ADS voice recognition models of Swiss-German (N=10) mothers separately were tested on two varieties of spontaneous speech 1) free-speech, and 2) picture-naming. The experiment was performed over 5 folds of cross-validation.

## DISCUSSION

Voice recognition is a process that is essential for social cohesion of humans and many other voice-recognizing animals, and behaviorally has been shown in infants both before and shortly after birth [46,47]. Here, we revealed that the universal phenomenon of IDS typically produced by caregivers to address their infants is a register that supports voice recognition performance when voices are learned under IDS. Using rigorous acoustic modelling we demonstrated – for the first time – that voice-specific timbre clusters capture more of the possible acoustic space in IDS compared to ADS, leading to better speaker recognition when caregivers speak in IDS. As a result, the acoustics of IDS provide optimal training input for the development of a robust voice identification system, one which has adapted to the tremendous situational variability in a caregiver's voice that the infant will encounter. It is thus likely that IDS is an adaptation applied by parents in infancy to foster acquisition of voice timbre detail by infants to subsequently support robust recognition of the caregiver by the infant, thus providing a basis for close bonding and increased chances of survival.

Our results are in line with recent findings from both face[35–37 35–37] and voice recognition [22], [37–40,37–40] revealing that viewing a face from different perspectives or listening to a voice of a speaker under increased variability supports subsequent face or voice recognition, respectively. This is in line with results showing that rich multi-modal information results in better representations of affective categories [48]. Similar evidence also exists from automatic voice recognition, demonstrating that training data from speakers under a variety of speech registers increases correct recognition probability of test data containing high variability [18]. This means that maximizing knowledge about within-voice variability supports the receivers' ability to identify a voice under a wide variety of registers and generalizes to previously unexperienced realizations of particular registers by a speaker. Here, we showed for the first time that the augmentation of within-speaker variability that is present in IDS fosters and supports enhanced speaker recognition as a result of a wide spectrum of within-

184 voice variability of the caregiver. Given a high-variability advantage detected for types of adult
185 directed speech [38], it seems plausible that IDS evolved for precisely such scenarios.

186 Within-speaker variability has typically been viewed as detrimental for recognition because the
187 detection of similarity in stimuli with high variability is more challenging [49,50]. However, our results
188 highlight the importance of variability as advantageous when supplied sufficiently during the training
189 period [38]. How long then does a familiarisation with IDS need to last to show advantages over speech
190 that consists of low-variability vocalisations? While the answer to this question is subject to future
191 studies, it seems plausible that after familiarization with IDS over the first year of life, the infant will
192 be well prepared to perform caregiver recognition by the time autonomous mobility (i.e., independent
193 crawling) starts and the self-exploration of different social settings consisting of new voices increases.
194 This is probably the point at which robust recognition of the primary caregiver is most crucial. Thus,
195 the point at which IDS changes into so-called child-directed speech (CDS) – which should then
196 contain less indexical signalling – may also be around that time. In fact, there is evidence for the
197 change between IDS and CDS to be indeed around that stage in the infant's life [51,52], [53,54].

198 Previous studies have predominantly focused on the hypothesis that the vocal adjustments in IDS
199 support language acquisition, in particular the process by which abstract linguistic forms (e.g.,
200 phonemes, syllables, or words) are acquired from variable acoustic representations [10]. However, the
201 type of timbre variability found in IDS introduces highly non-canonical linguistic realisations [55–57]
202 which have been found to produce ambiguous category information, thus, arguably,
203 counterproductive for the acquisition of linguistic categories [58] and shown to lead to less discriminable
204 utterances at the phoneme level [59]. A recent machine learning study [57] found that realisations of IDS
205 phoneme categories were neither more separable nor more robust in comparison to their ADS
206 counterparts. This is entirely consistent with the present results, as increasing voice timbre detail
207 should increase the variability found between segments, rendering them less canonical. Thus, it is
208 plausible that the dominant function of IDS is to elicit infant's attention to voice specific aspects.

Our results have wide-reaching implications. In evolutionary terms, it seems plausible that IDS has evolved as part of an adaptive set of strategies that serve to promote indexical signalling by an adult to their offspring, thereby promoting adult-infant bonding, increasing the infant's attention, and fostering offspring survival. Given the relatively poor linguistic skills of new-borns and infants in the first few months *post partum*, it seems plausible that the dominant role of voice in communication would be transmitting identity. This would also make offspring-directed vocalisations a plausible phenomenon in voice recognizing animals, where caregiver recognition may be similarly important. In fact, Fernandez and Knörnschild [60] recently provided evidence for the existence of an analogue of IDS in bats, which were shown to adopt a more highly variable vocal register when directing their calls to pups than adult conspecifics. Nevertheless, Fernandez and Knörnschild [60] automatically classified individual caregiver signals und found low classification rates. They concluded that pup-directed calls "do not seem to encode sufficient interindividual variation to allow for reliable individual discrimination" (p. 5). This, however, is not surprising, given that the advantage of higher variability signals only arises after long-term training [49,50]. The pup-directed vocalizations reported in [60] from 13 female pups – which evidently contain a high timbre variability – must be expected to result in poor classification performance and this result directly replicates findings for human speech [61] and our results from the within register tests – ADS-ADS and IDS-IDS – in which the higher variability of IDS has disadvantages over ADS recognition (see above). Hence, a lower acoustic variability in adult-directed bat vocalizations would mean that their classification rates should be higher compared to pup-directed vocalizations (classification results not reported in [60]). Bat pups, however, are evidently good at identifying their caregivers [62] and in the light of the results reported here, it is plausible that this performance is supported by the higher variability vocalizations reported in [60]. We predict that using pup-directed bat vocalizations as training material for a recognition task would show similar training advantages to the human data reported here. It will be interesting to see whether offspring-directed adjustments in vocalisations occur in other vocal recognizing animals (e.g. [63–66]) and whether these vocal adjustments co-occur with individual recognition advantages. Given that both human and

non-human animals share this non-linguistic function of vocalisations, it is plausible that vocalising to bootstrap identification may be the ancestral function of infant-directed vocalisations which was later leveraged to aid the development and realisation of linguistic contrasts in human speech [67,68].

**ACKNOWLEDGEMENTS**

Andrea Fröhlich and Sarah Lim carried out the recordings of the Swiss speakers and Marco Bleiker carrier out subject recruitment. Meisam K. Arjmandi assisted with implementing analyses on the US English dataset at Michigan State University. We also acknowledge Tonya Bergeson, Derek Houston, and Maria Kondaurova for assistance with the original US English dataset collection at the Devault Otologic Research Lab. This research was supported by grant #100012_125399 and grant NCCR Evolving Language #51NF40_180888 of the Swiss National Science Foundation.

# STAR METHODS

## Mothers Reveal More of Their Vocal Identities When Talking to Infants

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Software and Algorithms | | |
| MATLAB R2019b | Mathworks | RRID: SCR_001622 |
| Voicebox Speech processing toolbox | [Brookes, 2011] | http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html |
| HTK MFCC toolbox | [Davis and Mermelstein, 1980; Young et al., 2006] | https://ch.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab |
| R Studio | [R Core Team, 2013] | http://www.R-project.org/ |
| ggplot2 | [Wickham H, 2016] | https://ggplot2.tidyverse.org. |

### CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the corresponding author, Volker Dellwo (volker.dellwo@uzh.ch).

### EXPERIMENTAL MODEL AND SUBJECT DETAILS

*US-English materials:* 27 mother–infant dyads were recruited from the local community in Indianapolis, Indiana, USA; all mothers were native US English speakers with self-reported normal hearing who grew up in the Midwestern United States. Mothers participated in the study with their infants. Recordings were randomly selected from among those available such that infants were on average 7.5 months old (range: 4.1 - 9.9 months). The gender of infants was balanced. This research and the recruitment of human subjects were approved by the Indiana University Institutional Review Board. During a lab visit, mothers were asked to read for both IDS and ADS conditions from a storybook. In the IDS speech condition, mothers read the storybook to their infants, and the infants were present in the room with their mothers for the duration of the IDS session. In the ADS condition, mothers were asked to read aloud as if to another adult. The order of IDS and ADS recordings was counterbalanced across mothers and visits. For additional details, see [44]

*Swiss-German materials:* 10 Swiss-German mother–infant dyads participated in the experiment. All the mothers were native Swiss-German speakers with self-reported normal hearing. Infants were 8-9 months old, and the gender was balanced. Informed consent was obtained from all participating mothers, and approval of the study was obtained from the University of Zurich. Participants were told that we were broadly interested in ''mother-child interaction'' and were not aware that we were measuring differences between the acoustic properties of their speech across the conditions. The participants read 12 sentences to their infant, and to an adult experimenter in their own reading register.

## METHOD DETAILS

### Equipment

*US-English Recording*: Mothers were digitally recorded reading in a double-walled, copper-shielded sound booth (Industrial Acoustics Company, Bronx, NY) at the DeVault Otologic Research Lab in Indianapolis, Indiana, USA. The speech was recorded in one of two ways: The initial system used an Audio-Technica ES933/H hypercardioid microphone (Audio-Technica, Leeds, UK) powered by a phantom power source and linked to an amplifier (DSC-240; Daqscribe, Centennial, CO) and a Sony DTC-690 digital audio tape recorder (Sony, Tokyo, Japan). The equipment was updated partway through this longitudinal project to an SLX Wireless Microphone System (Shure, Niles, IL). This system included an SLX1 Bodypack transmitter with a built-in microphone and a wireless receiver SLX4, which was connected to a Canon 3CCD Digital Video Camcorder GL2, NTSC (Canon,Melville, NY) and recorded the speech samples directly onto a Mac computer (OSX Version 10.4.10; Apple, Inc., Cupertino, CA) via Hack TV (Version 1.11) software. No systematic differences were found across recording sessions or participant groups in terms of recording technology. Recordings were made at a sampling rate of 22050 Hz with 16-bit quantization rate. See [44] for more details.

*Swiss-German Recording*: Speech data were recorded continuously using Sennheiser (model: MKE 2 P) omni-directional prepolarized condensor clip-on microphone. The microphone was attached to

each mother's shirt collar. The sound recordings were done in a sound-treated room in the psychology department's baby lab led by author 6 at University of Zurich. Recordings were made at a sampling rate of 48 kHz Hz with 16-bit quantization rate. To obtain the same reading data for mothers and infants we recorded a small passage from a child-book in both the IDS and ADS conditions.

**Quantification and Statistical Analysis**

Within-speaker vocal variability can occur at different levels (e.g., phonemic, acoustic feature, speaking register; cf. [38]), and there is evidence for top-down recognition advantages through multiple levels of linguistic and paralinguistic features contributing to the voice recognition process [71]. Here we limit voice recognition to the process of recognizing a speaker based predominantly on the bottom-up acoustic properties of voice, i.e., independent of lexical choices or other language related properties. This is consistent with infants' limited knowledge of the higher-order linguistic properties of the environmental language, which requires that they must rely predominantly on acoustic characteristics of voice alone to recognize their caregivers.

**Feature Extraction**

MFCCs were obtained by block through processing the speech segment using a 25 ms Hamming window with an overlap of 10 ms, along with parameters that included pre-emphasis coefficient (0.95), number of filter bank channels (20), and liftering parameter (22), spectral bandwidth (150 - 8000 Hz) using MATLAB [Young et al, 2011]. The $0^{th}$ order coefficient (energy coefficient) were included in the 13-dimension MFCC features. The first- and second-order temporal derivatives were computed from the extracted MFCCs. Finally, 39 dimensional features (13 MFCCs, 13 first-order derivatives, and 13 second-order derivatives) per frame were used to develop a speaker model by training the Gaussian mixture models discussed in the ASR section.

**Number of Clusters in the GMM**

To compare the data spread of IDS compared to ADS within a cluster, we needed an optimal situation in which all clusters are filled with data from each category. Empty clusters were defined as clusters containing less than 4 data points in either IDS or ADS, corresponding to the minimum number of

points to calculate the data spread in 2-dimensional space, below. We performed a *k*-means cluster analysis with varying cluster numbers (16, 32, 64, 128, 256, 512); we found that from 64 clusters upwards, empty clusters start occurring. Interestingly, the number of empty clusters in ADS was always higher than in IDS. Given this finding, we used the 32 cluster analyses which reflected the point where empty clusters were equal between the two categories; we then measured several cluster properties: (a) distance between cluster means, (b) weight and (c) covariance.

**Automatic Speaker Recognition (ASR)**

We used GMM with diagonal covariance matrices [Reynolds et al., 2000] to design supervised language-independent speaker recognition models. In general, for GMM-based automatic speaker recognition experiments, the optimal number of Gaussian mixtures were estimated by varying the count ($2^N$; N = 1, 2, 3, 4...) until the best performance was obtained, which turned out to be 32 (N = 5) for most of the experiments. The speaker recognition model was designed for every speaker on both speaking registers separately, i.e., IDS and ADS per speaker. The extracted MFCCs with delta and delta-delta (for both, US-English and Swiss German dataset separately) were randomly split into an 80-20 ratio for training and testing. The data split and the ASR experiment was repeated using 5-fold cross-validation.

**Indexical Feature Expansion**

The 13-dimensional MFCC features (without delta and delta-delta) were normalized by speaker (*z*-score normalization within speakers). The normalized features were clustered using a *k*-means vector quantization algorithm [Voicebox tool in MATLAB, Brookes, 2011]. The cluster count was varied (shown in Figure S1. and Figure S2) to find the highest number of clusters (k) for which a minimal number of empty clusters (i.e. clusters with only one data sample) occur. To measure the data spread of a cluster, it is necessary to have more than four data samples in a cluster [Stat_Ellipse() function in ggplot2 R package]. At $k = 32$ clusters, there were 1 and 3 empty clusters found in Swiss-German and US-English datasets respectively. Solutions for $k > 32$ had notably more empty clusters, hence our choice was to use 32 clusters for subsequent analyses. After clustering, every data frame was assigned

a cluster label (1 - 32) and the speech register label (IDS or ADS). The 13-dimensional MFCC feature space was reduced to 2 dimension using principal component analysis (PCA). The surface area of IDS and ADS points in each cluster was calculated in terms of their ellipse surface area ($\pi ab$) in the 2-dimensional feature space (normal distribution assumed) and was used as a measure of timbre expansion in each sub-cluster.

**Control Analysis: Behavioural Classification of IDS Versus ADS**

In the Swiss data retrieval, babies could often not be separated from their mother during the adult recording. This resulted in vocalisations that were audibly less well separated on an IDS-ADS register continuum. To make sure that individual productions of IDS and ADS were categorizable, we tested the sensitivity (A') of 30 native listeners of Swiss German to identify individual productions as either IDS or ADS. The stimulus set consisted of 12 sentences read by 4 speakers in 2 registers ($N = 240$). This set was randomly divided into 3 subsets of N=80 each (4 sentences * 10 speakers * 2 registers). Each 10 randomly selected listeners listened to one of the three stimulus sets. The average A' revealed a high sensitivity of 0.79 and no listener bias (B''D = 0.1).

**Control Analysis: Automatic Classification of IDS Versus ADS**

To ensure that the data is comparable to previous studies, we replicated the results from Piazza et al. (2017) using a support-vector machine classifier (SVM-RBF) [MATLAB SVM classifier Application] for IDS-ADS classification based on 12 MFCCs (frames were not averaged) with a training/test split of 90/10 (5-fold cross-validation) on two datasets. Piazza et al. trained a linear SVM with ADS and found that speaker classification performance was decreased for IDS vs ADS. Based on this finding, they conclude that there is a systematic shift in timbral properties of the two registers. Our data replicates this finding with 70.3% and 75.8% correct classification rates for US English and Swiss-German respectively, consistent with the average classification of ~70% in [4] for English and non-English data. Importantly, failure of machine classification alone does not necessarily reflect a systematic acoustic timbre difference (shift) between ADS and IDS.

**Supplementary Text**

All the ASR experiments (Figure S3 and S4) have shown a similar pattern in the results i.e., the recognition accuracy of the model trained on IDS and tested with ADS is always better than the recognition accuracy of the model trained on ADS and tested on IDS independent of the features used, and the number of GMM mixtures varied. Concerning features, MFCCs with delta have shown a better mean accuracy than MFCCs alone in both datasets. The Swiss-German dataset revealed better mean accuracy than the US-English dataset. Overall, the recognition accuracy of the two datasets has declined with increase in GMM mixtures except for the MFCCs with delta in the Swiss-German dataset (Figure S4 b). In Figure S4 b, the peak performance was at 128 GMM mixtures but, for other experiments, the peak performance was at 32 GMM mixtures.

## Figures S1-S6

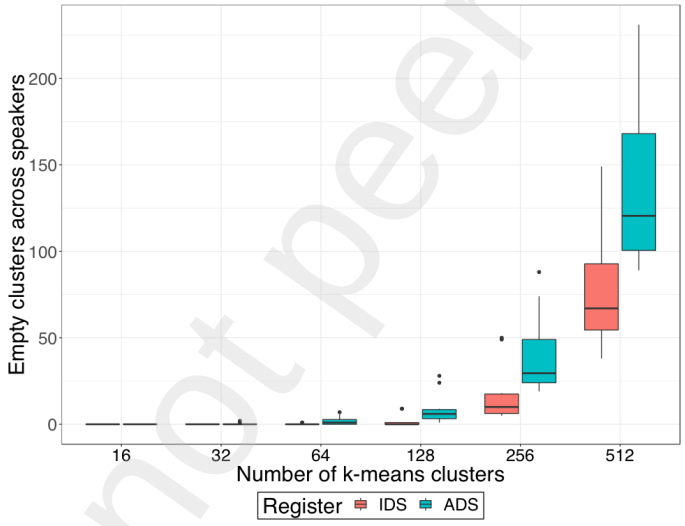
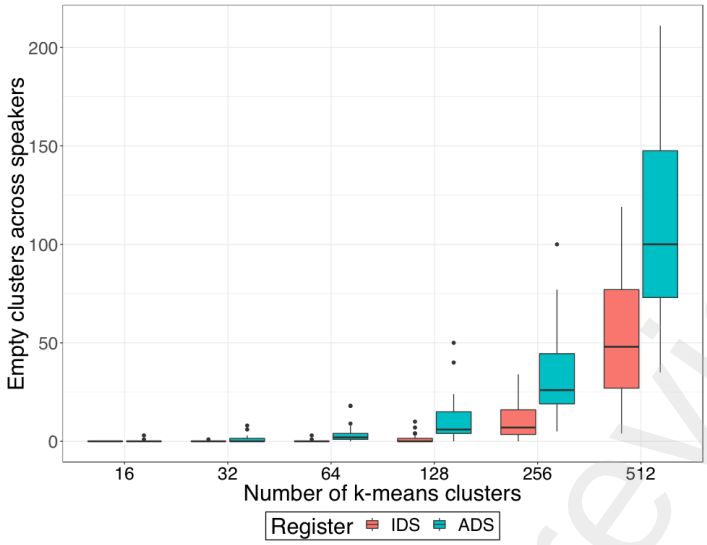


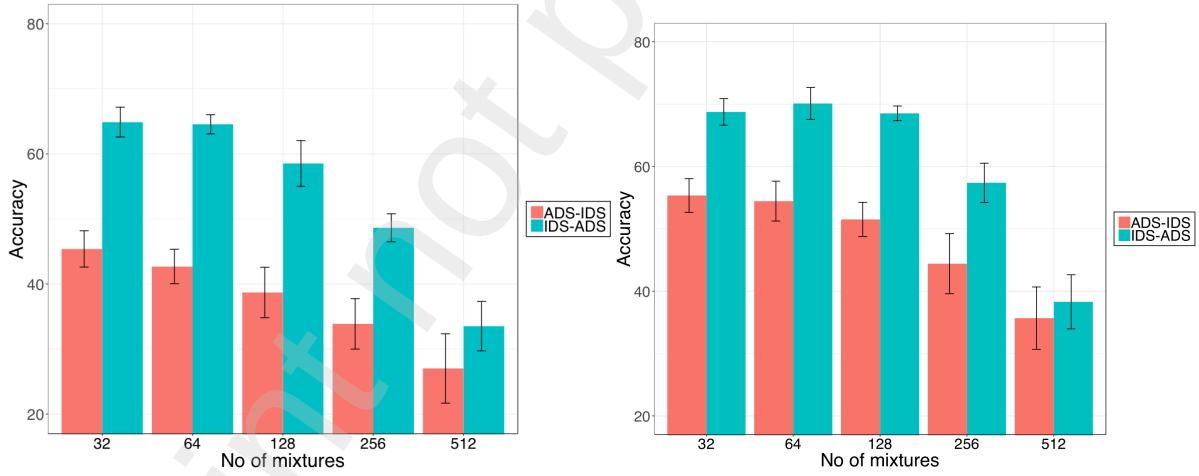

**Figure S1.** Empty-clusters in both IDS and ADS registers of 10 Swiss-German mothers in *k*-means clustering. The *x*-axis shows the number of *k*-means clusters used in the analysis, and the *y*-axis shows the variation of empty-clusters across 10 speakers. The vertical bar in the graph shows the standard deviation.




382

383 **Figure S2.** The occurrence of empty-clusters in both IDS and ADS registers of 27 US-English mothers
384 in *k*-means clustering. The *x*-axis shows the number of *k*-means clusters used in the analysis, and the *y*-
385 axis shows the variation of empty-clusters across 27 speakers. The vertical bar in the graph shows the
386 standard deviation.

387 A) B)

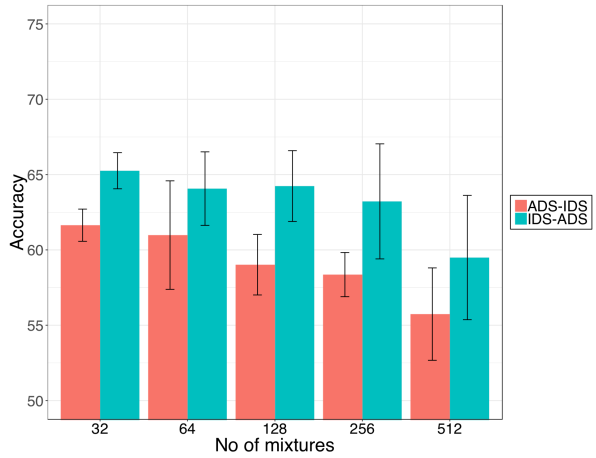



388

389 **Figure S3.** Cross-register voice recognition accuracy (mean and standard deviation) for US-English
390 dataset using a) MFCCs, and b) MFCCs with delta. The *x*-axis shows the number of mixtures in the
391 GMM classifier, and the *y*-axis shows the recognition accuracy in %. The vertical bar in the graph shows
392 the standard deviation of 5-fold cross validation.

393

A)




B)

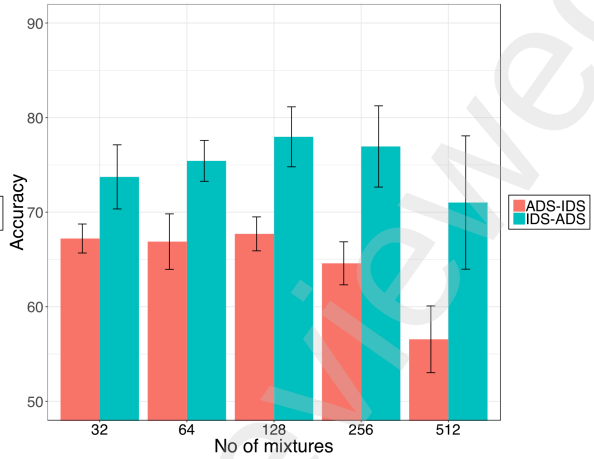



**Figure S4.** Cross-register voice recognition accuracy on Swiss-German dataset using a) MFCCs, and b) MFCCs with delta. The number of mixtures in the GMM classifier varied from 32 to 512, in the power of 2. The X-axis shows the number of mixtures in the GMM classifier, and the Y-axis shows the recognition accuracy in %, the vertical bar in the graph shows the standard-deviation of 5-fold cross validation.

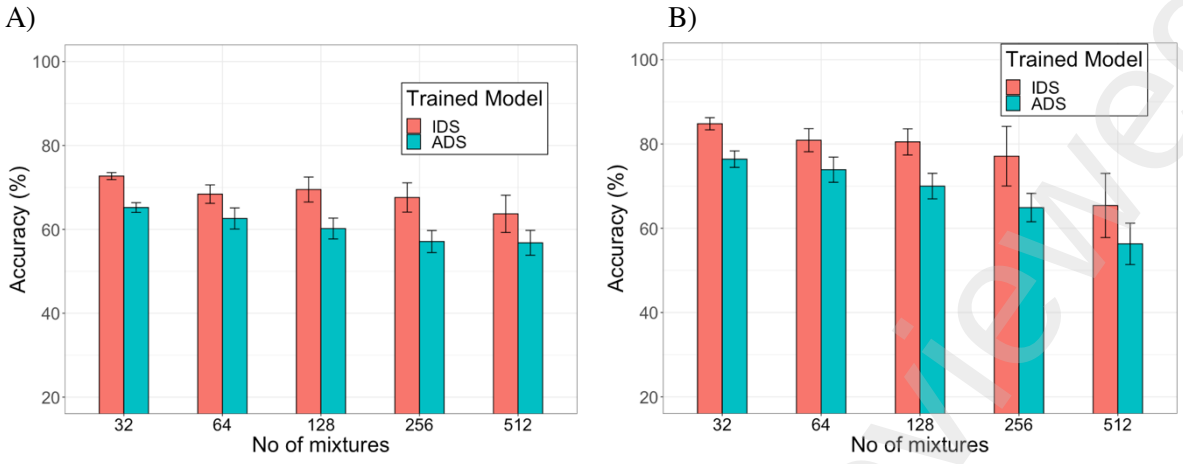



**Figure S5.** Speaker recognition accuracy on spontaneous speech (picture naming task in Swiss-German dataset) using a) MFCCs, and b) MFCCs with delta. The number of mixtures in the GMM classifier varied from 32 to 512, in the power of 2. The *x*-axis shows the number of mixtures in the GMM classifier, and the *y*-axis shows the recognition accuracy in %. The vertical bar in the graph shows the standard deviation of 5-fold cross validation.

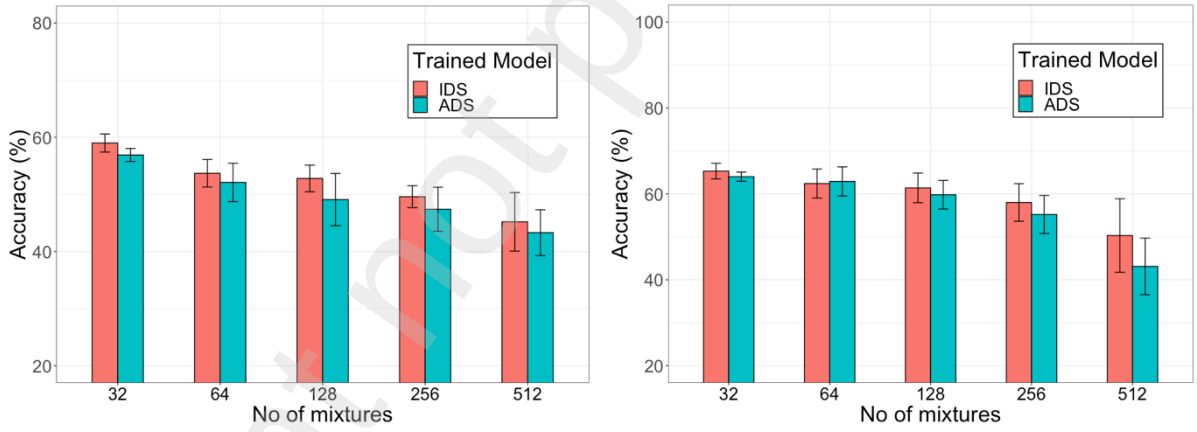



**Figure S6.** Voice recognition accuracy on spontaneous speech (free speech in Swiss-German dataset) using a) MFCCs, and b) MFCCs with delta. The number of mixtures in the GMM classifier varied from 32 to 512, in the power of 2. The *x*-axis shows the number of mixtures in the GMM classifier, and the *y*-axis shows the recognition accuracy in %. The vertical bar in the graph shows the standard-deviation of 5-fold cross validation.

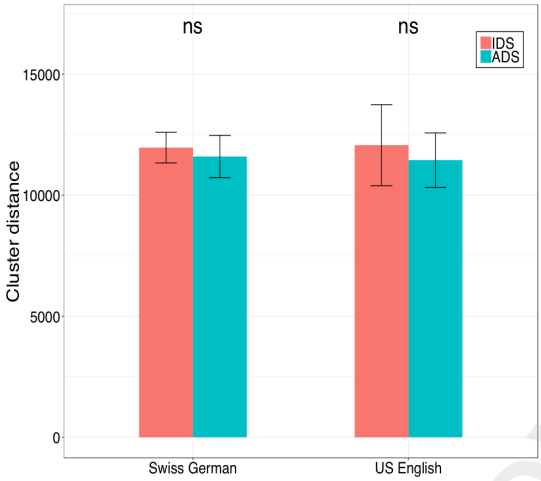



**Figure S7.** Sum of distance between centroids of the 32 IDS and ADS registers of Swiss German and US English dataset.

We also obtained the sum of the distances between cluster centroids in the IDS and the ADS feature space. Fig. S7 shows that the overall distance between IDS sub-clusters was higher compared to ADS sub-clusters (nonsignificant interaction between register and language: $F[1,35] = 0.3612$; $p = 0.5517$; significant main effect of registers: $F[1,35] = 5.6817, p < 0.0227$, but not of language: $F[1,35] = 0.0030, p < 0.9566$), meaning that IDS clusters are on average more widely dispersed compared to ADS clusters.

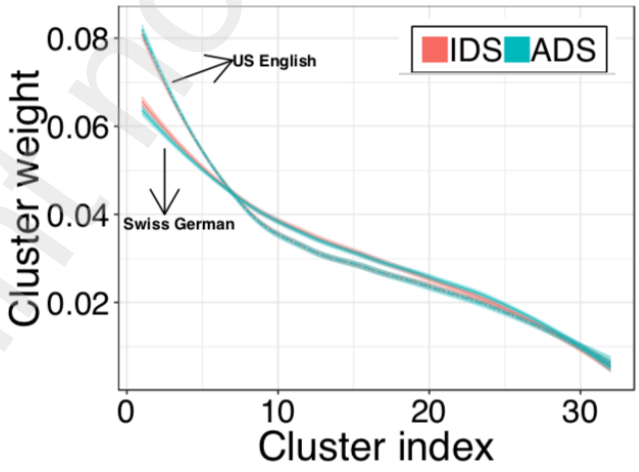



**Figure S8.** Cluster weight distribution of 32 clusters in IDS and ADS registers of the Swiss German and US English datasets.

## REFERENCES


1. Kisilevsky, B.S., Hains, S.M.J., Lee, K., Xie, X., Huang, H., Ye, H.H., Zhang, K., and Wang, Z. (2003). Effects of experience on fetal voice recognition. Psychological Science *14*, 220–224.
2. Beauchemin, M., González-Frankenberger, B., Tremblay, J., Vannasing, P., Martínez-Montes, E., Belin, P., Béland, R., Francoeur, D., Carceller, A.M., Wallois, F., et al. (2011). Mother and stranger: An electrophysiological study of voice processing in newborns. Cerebral Cortex *21*, 1705–1711.
3. Johnson, E.K., Westrek, E., Nazzi, T., and Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. Developmental Science *14*, 1002–1011.
4. Piazza, E.A., Iordan, M.C., and Lew-Williams, C. (2017). Mothers Consistently Alter Their Unique Vocal Fingerprints When Communicating with Infants. Current Biology *27*, 3162-3167.e3.
5. Webb, A.R., Heller, H.T., Benson, C.B., and Lahav, A. (2015). Mother's voice and heartbeat sounds elicit auditory plasticity in the human brain before full gestation. Proc Natl Acad Sci U S A *112*, 3152–3157.
6. Blasi, A., Mercure, E., Lloyd-Fox, S., Thomson, A., Brammer, M., Sauter, D., Deeley, Q., Barker, G.J., Renvall, V., Deoni, S., et al. (2011). Early specialization for voice and emotion processing in the infant brain. Current Biology *21*, 1220–1224.
7. Moser, C.J., Lee-Rubin, H., Bainbridge, C.M., Atwood, S., Simson, J., Knox, D., Glowacki, L., Galbarczyk, A., Jasienska, G., Ross, C.T., et al. (2020). Acoustic regularities in infant-directed vocalizations across cultures. bioRxiv, 2020.04.09.032995.
8. Broesch, T.L., and Bryant, G.A. (2015). Prosody in Infant-Directed Speech Is Similar Across Western and Traditional Cultures. Journal of Cognition and Development *16*, 31–43.
9. Kendrick, K.M. (2006). Introduction. The neurobiology of social recognition, attraction and bonding. Philos Trans R Soc Lond B Biol Sci *361*, 2057–2059.
10. Kuhl, P.K., Andruski, J.E., Chistovich, I.A., Chistovich, L.A., Kozhevnikova, E. v., Ryskina, V.L., Stolyarova, E.I., Sundberg, U., and Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. Science (1979) *277*, 684–686.
11. Latinus, M., McAleer, P., Bestelmeyer, P.E.G., and Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. Current Biology *23*, 1075–1080.
12. Lavan, N., Burston, L.F.K., and Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. British Journal of Psychology *110*, 576–593.
13. Lee, Y., Keating, P., and Kreiman, J. (2019). Acoustic voice variation within and between speakers. J Acoust Soc Am *146*, 1568–1579.
14. Hansen, J.H.L.L., and Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. IEEE Signal Processing Magazine *32*, 74–99.
15. Smith, H.M.J., Baguley, T.S., Robson, J., Dunn, A.K., and Stacey, P.C. (2019). Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance. Applied Cognitive Psychology *33*, 272–287.

16. Ramírez López, A., Saeidi, R., Juvela, L., and Alku, P. (2017). Normal-to-shouted speech spectral mapping for speaker recognition under vocal effort mismatch. In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 4940–4944.
17. Jokinen, E., Saeidi, R., Kinnunen, T., and Alku, P. (2019). Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task. Computer Speech and Language *53*, 1–11.
18. Shriberg, E., Graciarena, M., Bratt, H., Kathol, A., Kajarekar, S., Jameel, H., Richey, C., and Goodman, F. (2008). Effects of vocal effort and speaking style on text-independent speaker verification. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 609–612.
19. Shriberg, E., Graciarena, M., Bratt, H., Kathol, A., International, S.R.I., and Park, M. (2008). Effects of Vocal Effort and Speaking Style on Text-Independent Speaker Verification. In INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, pp. 609–612.
20. Kuhl, P.K. (2004). Early language acquisition: cracking the speech code. Nature Reviews Neuroscience 2004 5:11 *5*, 831–843.
21. Kitamura, C., and Burnham, D. (2003). Pitch and Communicative Intent in Mother's Speech: Adjustments for Age and Sex in the First Year. Infancy *4*, 85–110.
22. Singh, L., Morgan, J.L., and Best, C.T. (2002). Infants' Listening Preferences: Baby Talk or Happy Talk? Infancy *3*, 365–394.
23. Golinkoff, R.M., Can, D.D., Soderstrom, M., and Hirsh-Pasek, K. (2015). (Baby)Talk to Me: The Social Context of Infant-Directed Speech and Its Effects on Early Language Acquisition. Current Directions in Psychological Science *24*, 339–344.
24. Gauthier, B., and Shi, R. (2011). A connectionist study on the role of pitch in infant-directed speech. J Acoust Soc Am *130*, EL380–EL386.
25. Kalashnikova, M., Carignan, C., and Burnham, D. (2017). The origins of babytalk : smiling , teaching or social convergence ? Subject Category : Subject Areas : Author for correspondence :
26. Burnham, D., Kitamura, C., and Vollmer-Conna, U. (2002). What's New, Pussycat? On Talking to Babies and Animals. Science (1979) *296*, 1435 LP – 1435.
27. Miyazawa, K., Shinya, T., Martin, A., Kikuchi, H., and Mazuka, R. (2017). Vowels in infant-directed speech: More breathy and more variable, but not clearer. Cognition *166*, 84–93.
28. van der Feest, S.V.H., Blanco, C.P., and Smiljanic, R. (2019). Influence of speaking style adaptations and semantic context on the time course of word recognition in quiet and in noise. Journal of Phonetics *73*, 158–177.
29. Cooper, R.P., and Aslin, R.N. (1989). The language environment of the young infant: Implications for early perceptual development. Canadian Journal of Psychology/Revue canadienne de psychologie *43*, 247–265.
30. Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., and Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. Journal of Child Language *16*, 477–501.
31. Fernald, A., and Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. Developmental Psychology *20*, 104–113.

32. Lee, C.S., Kitamura, C., Burnham, D., and McAngus Todd, N.P. (2014). On the rhythm of infant- versus adult-directed speech in Australian English. J Acoust Soc Am *136*, 357–365.
33. Payne, E.M., Post, B., Astruc, L., Prieto, P., and Vanrell, M. (2015). Rhythmic modification in child directed speech. In Supplementi alla biblioteca di linguistica., M. Russo, ed. (Aracne), pp. 147–183.
34. Leong, V., Kalashnikova, M., Burnham, D., and Goswami, U. (2017). The Temporal Modulation Structure of Infant-Directed Speech. Open Mind: Discoveries in Cognitive Science, 1–13.
35. Kramer, R.S.S., Young, A.W., and Burton, A.M. (2018). Understanding face familiarity. Cognition *172*, 46–58.
36. Kramer, R.S.S., Jenkins, R., Young, A.W., and Burton, A.M. (2017). Natural variability is essential to learning new faces. Visual Cognition *25*, 470–476.
37. Burton, A.M., Kramer, R.S.S., Ritchie, K.L., and Jenkins, R. (2016). Identity From Variation: Representations of Faces Derived From Multiple Instances. Cognitive Science *40*, 202–223.
38. Lavan, N., Knight, S., Hazan, V., and McGettigan, C. (2019). The effects of high variability training on voice identity learning. Cognition *193*, 104026.
39. Karlsson, I. (1999). Within-speaker variability in the VeriVox database. Fonetik '99: Proceedings from the Twelfth Swedish Phonetics Conference *81*, 93–96.
40. Karlsson, I., Banziger, T., Dankovicová, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., and Scherer, K. (2000). Speaker verification with elicited speaking styles in the VeriVox project. Speech Communication *31*, 121–129.
41. Davis, S.B., and Mermelstein, P. (1980). Comparison of Parametric Representations for. Transactions on Acoustics, Speech, and Signal Processing *28*, 357–366.
42. Friendly, M., Monette, G., and Fox, J. (2013). Elliptical insights: Understanding statistical methods through elliptical geometry. Statistical Science *28*, 1–39.
43. Wieland, E.A., Burnham, E.B., Kondaurova, M., Bergeson, T.R., and Dilley, L.C. (2015). Vowel space characteristics of speech directed to children with and without hearing loss. Journal of Speech, Language, and Hearing Research *58*, 254–267.
44. Dilley, L., Lehet, M., Wieland, E.A., Arjmandi, M.K., Kondaurova, M., Wang, Y., Reed, J., Svirsky, M., Houston, D., and Bergeson, T. (2020). Individual Differences in Mothers' Spontaneous Infant-Directed Speech Predict Language Attainment in Children With Cochlear Implants. Journal of Speech, Language, and Hearing Research *63*, 2453–2467.
45. Cristia, A. (2022). A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations. Developmental Science.
46. DeCasper, A.J., and Fifer, W.P. (1980). Of Human Bonding - Newborns Prefer Their Mothers Voices. Science (1979) *208*, 1174–1176.
47. Fecher, N., and Johnson, E.K. (2019). By 4.5 Months, Linguistic Experience Already Affects Infants' Talker Processing Abilities. Child Development *90*, 1535–1543.
48. Flom, R., and Bahrick, L.E. (2007). The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. Developmental Psychology *43*, 238–252.

49. Lavan, N., Knight, S., Hazan, V., and Mcgettigan, C. (2016). No clear advantage for high variability training during voice identity learning.
50. Lavan, N., Burton, A.M., Scott, S.K., and McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. Psychonomic Bulletin and Review *26*, 90–102.
51. Vosoughi, S., and Roy, D. (2012). A Longitudinal Study of Prosodic Exaggeration in Child-directed Speech. Proceedings of the 6th International Conference on Speech Prosody (SP2012), 194–197.
52. Liu, H.-M., Tsao, F.-M., and Kuhl, P.K. (2009). Age-related changes in acoustic modifications of Mandarin maternal speech to preverbal infants and five-year-old children: a longitudinal study. J. Child Lang. *36*, 909–922.
53. Ratner, N.B. (1984). Patterns of vowel modification in mother–child speech. Journal of Child Language *11*, 557–578.
54. Ko, E.-S. (2012). Nonlinear development of speaking rate in child-directed speech. Lingua *122*, 841–857.
55. Friedrichs, D., Maurer, D., Rosen, S., and Dellwo, V. (2017). Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks. J Acoust Soc Am *142*, 1025–1033.
56. Kathiresan, T., Maurer, D., and Dellwo, V. (2019). Highly spectrally undersampled vowels can be classified by machines without supervision. J Acoust Soc Am *146*, EL1–EL7.
57. Ludusan, B., Mazuka, R., and Dupoux, E. (2021). Does Infant-Directed Speech Help Phonetic Learning? A Machine Learning Investigation. Cognitive Science *45*, 12946.
58. McMurray, B., Kovack-Lesh, K.A., Goodwin, D., and McEchron, W. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? Cognition *129*, 362–378.
59. Guevara-Rukoz, A., Cristia, A., Ludusan, B., Thiollière, R., Martin, A., Mazuka, R., and Dupoux, E. (2018). Are Words Easier to Learn From Infant- Than Adult-Directed Speech? A Quantitative Corpus-Based Investigation. Cognitive Science *42*, 1586–1617.
60. Fernandez, A.A., and Knörnschild, M. (2020). Pup Directed Vocalizations of Adult Females and Males in a Vocal Learning Bat. Frontiers in Ecology and Evolution *8*.
61. Lavan, N., Knight, S., Hazan, V., and Mcgettigan, C. (2016). No clear advantage for high variability training during voice identity learning.
62. Balcombe, J.P., and McCracken, G.F. (1992). Vocal recognition in Mexican free-tailed bats: Do pups recognize mothers? Animal Behaviour *43*, 79–87.
63. Jouventin, P., and Aubin, T. (2002). Acoustic systems are adapted to breeding ecologies: individual recognition in nesting penguins. Animal Behaviour *64*, 747–757.
64. Insley, S.J. (2001). Mother-Offspring vocal recognition in northern fur seals is mutual but asymmetrical. Anim Behav *61*, 129–137.
65. de Fanis, E., and Jones, G. (1995). Post-natal growth, mother-infant interactions and development of vocalizations in the vespertilionid bat Plecotus auritus. Journal of Zoology *235*, 85–97.

66. Torriani, M.V.G., Vannoni, E., and McElligott, A.G. (2006). Mother-Young Recognition in an Ungulate Hider Species: A Unidirectional Process. The American Naturalist *168*, 412–420.
67. Creel, S.C., and Bregman, M.R. (2011). How Talker Identity Relates to Language Processing. Linguistics and Language Compass *5*, 190–204.
68. Belin, P. (2006). Voice processing in human and non-human primates. Philosophical Transactions of the Royal Society B: Biological Sciences *361*, 2091–2107.
69. Lee, G.Y., and Kisilevsky, B.S. (2014). Fetuses respond to father's voice but prefer mother's voice after birth. Developmental Psychobiology *56*, 1–11.
70. Ward, C.D., and Cooper, R.P. (1999). A lack of evidence in 4-month-old human infants for paternal voice preference. Developmental Psychobiology *35*, 49–59.
71. Zarate, J.M., Tian, X., Woods, K.J.P., and Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. Scientific Reports *5*, 11475.