Current Biology

Report

Mothers Consistently Alter Their Unique Vocal Fingerprints When Communicating with Infants

Highlights

- Infant-directed speech is an important mode of communication for early learning
- Mothers shift the statistics of their vocal timbre when speaking to infants
- This systematic shift generalizes robustly across a variety of languages
- This research has implications for infant learning and speech recognition technology

Authors

Elise A. Piazza, Marius Cătălin Iordan, Casey Lew-Williams

Correspondence

elise.piazza@gmail.com

In Brief

Piazza et al. report a novel feature of motherese. When communicating with their infants, mothers shift the summary statistics of their vocal spectra, thereby altering their unique timbre fingerprints. This shift generalizes across a wide variety of languages and thus may be a universal form of communication with infants.

Piazza et al., 2017, Current Biology *27*, 1–6
October 23, 2017 © 2017 Elsevier Ltd.
https://doi.org/10.1016/j.cub.2017.08.074

CellPress

Please cite this article in press as: Piazza et al., Mothers Consistently Alter Their Unique Vocal Fingerprints When Communicating with Infants, Current Biology (2017), https://doi.org/10.1016/j.cub.2017.08.074

Current Biology
Report

Mothers Consistently Alter Their Unique Vocal Fingerprints When Communicating with Infants

Elise A. Piazza,[1,2,3,*] Marius Cătălin Iordan,[1,2] and Casey Lew-Williams[2]
[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
[2]Department of Psychology, Princeton University, Princeton, NJ 08544, USA
[3]Lead Contact
*Correspondence: elise.piazza@gmail.com

https://doi.org/10.1016/j.cub.2017.08.074

SUMMARY

The voice is the most direct link we have to others' minds, allowing us to communicate using a rich variety of speech cues [1, 2]. This link is particularly critical early in life as parents draw infants into the structure of their environment using infant-directed speech (IDS), a communicative code with unique pitch and rhythmic characteristics relative to adult-directed speech (ADS) [3, 4]. To begin breaking into language, infants must discern subtle statistical differences about people and voices in order to direct their attention toward the most relevant signals. Here, we uncover a new defining feature of IDS: mothers significantly alter statistical properties of vocal timbre when speaking to their infants. Timbre, the tone color or unique quality of a sound, is a spectral fingerprint that helps us instantly identify and classify sound sources, such as individual people and musical instruments [5–7]. We recorded 24 mothers' naturalistic speech while they interacted with their infants and with adult experimenters in their native language. Half of the participants were English speakers, and half were not. Using a support vector machine classifier, we found that mothers consistently shifted their timbre between ADS and IDS. Importantly, this shift was similar across languages, suggesting that such alterations of timbre may be universal. These findings have theoretical implications for understanding how infants tune in to their local communicative environments. Moreover, our classification algorithm for identifying infant-directed timbre has direct translational implications for speech recognition technology.

RESULTS

If mothers systematically alter their unique timbre signatures when speaking to their infants, we predicted that we could use freely improvised, naturalistic speech data to discriminate infant-directed from adult-directed speech. Furthermore, if this systematic shift in timbre production during IDS exists, we expected that it would manifest similarly across a wide variety of languages.

Twenty-four mother-infant dyads participated in this study (see STAR Methods). We recorded mothers' naturalistic speech while they spoke to their infants and to an adult interviewer. During the recorded session, half of the mothers spoke only English and the other half spoke only a non-English language (the language they predominantly used when speaking to their child at home). For each participant, we extracted 20 short utterances from each condition (IDS, ADS) and computed a single, time-averaged MFCC vector (i.e., a concise summary statistic representing the signature tone "color" of that mother's voice; see Figure 1 and STAR Methods) from each utterance. This measure—a limited set of time-averaged values that concisely describe a sound's unique spectral properties [8, 9]—has been shown to represent human timbre perception quite well [7]. As an initial validation of our method, we first confirmed that support-vector machine (SVM) classification is sensitive enough to replicate previous work distinguishing individual mothers [10, 11] by performing the classification on these MFCC vectors across subjects (see STAR Methods). Then, to test our primary question of interest, we performed a similar SVM classification on these vectors to distinguish IDS from ADS. Our use of MFCC as a global summary measure of vocal signature across varied, naturalistic speech (see STAR Methods) represents a new approach to discriminating communicative modes in real-life contexts.

Classification of Infant- versus Adult-Directed Speech

Using a support-vector machine classifier (SVM-RBF, see STAR Methods), we were able to distinguish utterances of infant-directed from adult-directed speech significantly above chance using the MFCC, here used as a summary statistical feature vector that represents the overall timbre fingerprint of someone's voice (see STAR Methods). Our classification analysis discriminated IDS from ADS for both English speech (Figure 2; two-tailed, one-sample t test, $t(11) = 6.85$, $p < 0.0001$) and non-English speech ($t(11) = 4.84$, $p < 0.001$). These results indicate that timbre shifts across communicative modes are highly consistent across mothers.

Classification of Infant- versus Adult-Directed Speech Generalizes across Languages

In a cross-language decoding analysis, we also found that the classifier trained to distinguish English IDS from ADS could

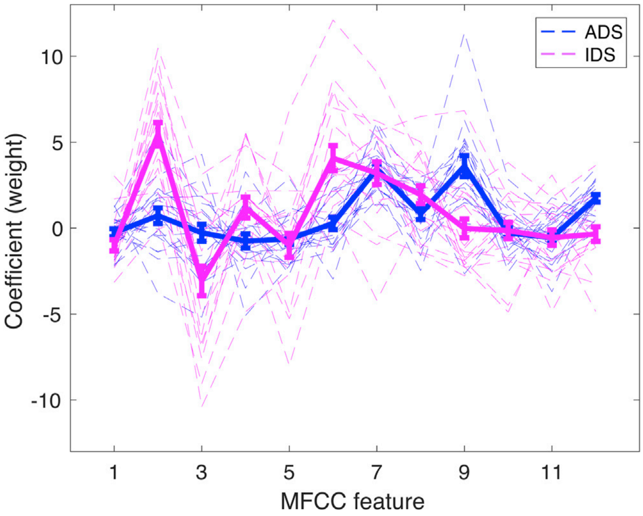

Figure 1. MFCC Feature Vectors from All Utterances for One Representative Participant

Each vector (dashed lines) represents the time-averaged set of Mel-frequency cepstral coefficients for a single utterance of either adult-directed speech (ADS, shown in blue) or infant-directed speech (IDS, shown in pink). Each bold line represents the average MFCC vector across all 20 utterances for a given condition. Error bars on the averaged vectors represent ±SEM across 20 utterances. Figure S1 depicts average MFCC vectors for each of the 12 English-speaking participants; the vectors displayed in this figure come from s12.

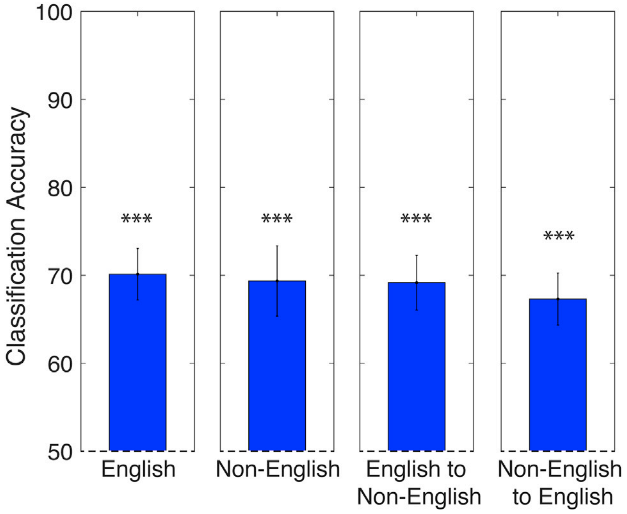

Figure 2. Accuracy Rates for Classifying Mothers' IDS versus ADS using MFCC Vectors

The first two bars indicate results from training and testing the classifier on English (first bar) and on all other languages (second bar). The third bar results from training the classifier on English data and testing on non-English data (and vice versa for the fourth bar). Chance (dashed line) is 50%. N = 12. Classification performance is represented as mean percent correct and ±SEM across cross-validation folds (leave-one-subject-out). ***p < 0.001.

discriminate these two modes of speech significantly above chance when tested instead on non-English data (Figure 2, $t(11) = 6.16$, $p < 0.0001$). Conversely, a classifier trained to distinguish non-English IDS from ADS could also successfully discriminate English data ($t(11) = 5.84$, $p < 0.001$). Thus, the timbral transformation used (perhaps automatically) by English speakers when switching from ADS to IDS generalizes robustly to other languages. See STAR Methods for the full list of languages tested.

Classification of Infant- versus Adult-Directed Speech Cannot Be Fully Explained by Differences in Pitch (F0) or Formants (F1, F2)

We performed a control analysis to rule out the possibility that our ability to distinguish IDS from ADS was based on differences in pitch that were somehow recoverable in the MFCC feature set. To do this, we regressed out F0 (over time) from each of the 12 MFCC time vectors before computing the single time-averaged vector of MFCC coefficients for each utterance and performing classification based on those time-averaged vectors (see STAR Methods). After removing the dynamic effects of F0, we found that classification between ADS and IDS (using the same algorithm as in previous analyses, SVM-RBF) was still significantly above chance for English data (Figure 3A, second bar; two-tailed, one-sample t test, $t(11) = 4.45$, $p < 0.001$), non-English data (Figure 3B, second bar; $t(11) = 4.61$, $p < 0.001$), training on English and testing on non-English data (Figure 3C, second bar; $t(11) = 5.54$, $p < 0.001$), and training on non-English and testing on English data (Figure 3D, second bar; $t(11) = 5.36$, $p < 0.001$). Thus, even in the absence of pitch differences, timbre information alone enabled robust discrimination of ADS and IDS.

In a second analysis, we computed the F1 and F2 contours from each utterance and regressed out each of these vectors (in addition to F0) from each MFCC time vector before computing the single time-averaged vector of MFCC coefficients for each utterance. We chose F1 and F2 because they have been identified in previous research as properties of speech that are shifted between IDS and ADS [12]. After removing the dynamic effects of F0 and the first two formants, we found that classification was still significantly above chance for English data (Figure 3A, third bar; two-tailed, one-sample t test, $t(11) = 3.88$, $p < 0.01$), non-English data (Figure 3B, third bar; $t(11) = 4.85$, $p < 0.001$), training on English and testing on non-English data (Figure 3C, third bar; $t(11) = 8.31$, $p < 0.0001$), and training on non-English and testing on English data (Figure 3D, third bar; $t(11) = 3.37$, $p < 0.01$). Thus, even after removing the dynamic effects of pitch and the first two formants, the remaining timbre information present in MFCCs enabled robust discrimination of ADS and IDS.

Classification of IDS versus ADS Data Cannot Be Fully Explained by Differences in Background Noise

We performed another control analysis to rule out the possibility that our ability to distinguish between IDS and ADS was due to differences between the noise properties of the microphone or room across different conditions. We were able to classify ADS versus IDS using silence alone for English speakers (Figure 4A, yellow bar; two-tailed, one-sample t test, $t(11) = 3.26$, $p < 0.01$) and for non-English speakers (Figure 4B, yellow bar; $t(11) = 2.33$, $p < 0.05$). This could result from slight shifts in microphone position when mothers were oriented toward the adult experimenter versus their infant. But importantly, classification of ADS versus IDS was significantly better for real speech than

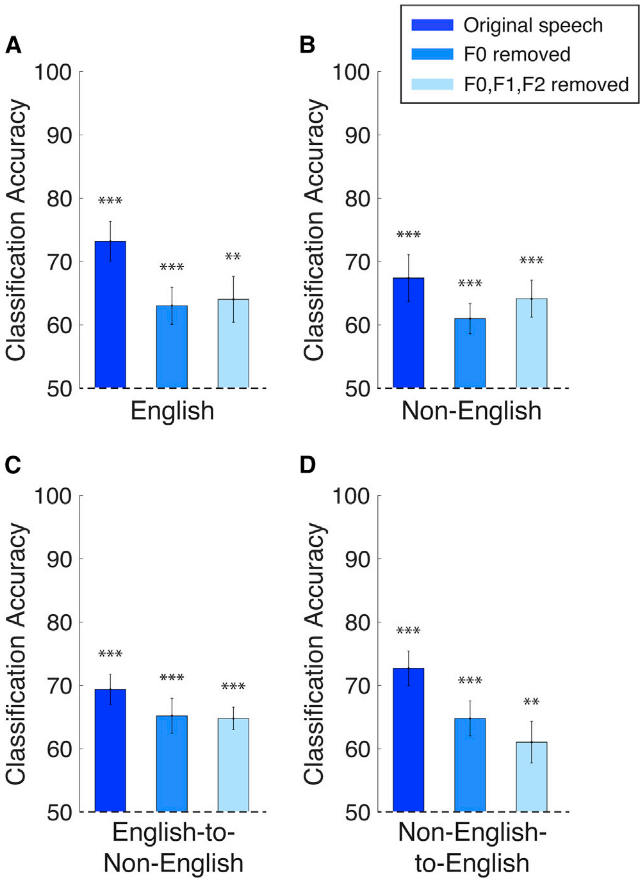

Figure 3. Accuracy Rates for Classifying IDS versus ADS Based on Timbre, after Controlling for Pitch and Formants

The first (darkest blue) bars indicate results for MFCC vectors derived from original speech, the second bars indicate results from speech with F0 regressed out, and the third bars indicate results from speech with F0, F1, and F2 regressed out. Bars corresponding to "original speech" are derived from only the segments of each utterance in which an F0 value was obtained, for direct comparison with the regression results (see STAR Methods). Chance (dashed line) is 50%. N = 12. Classification performance is represented as mean percent correct and ±SEM across cross-validation folds (leave-one-subject-out). **p < 0.01, ***p < 0.001.

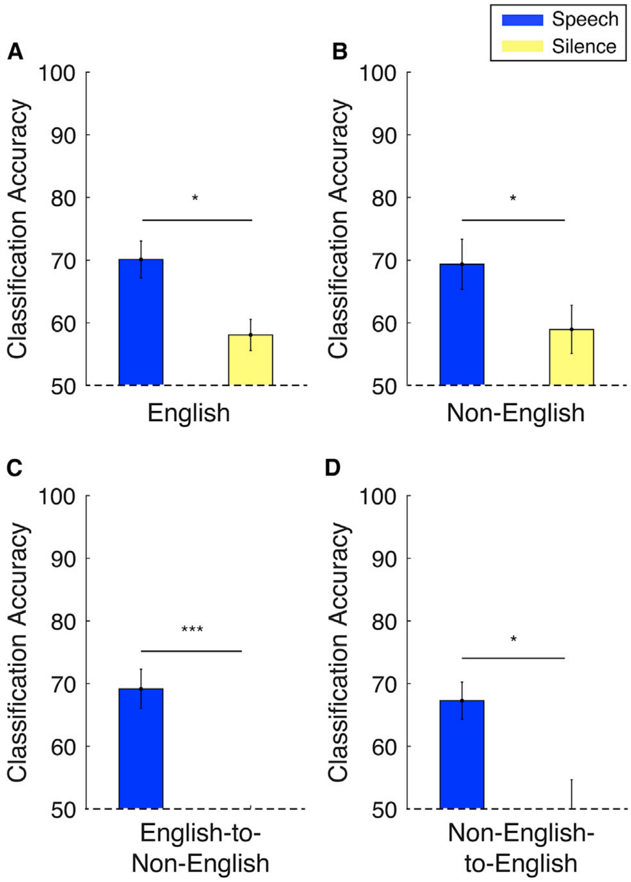

Figure 4. Accuracy Rates for Classifying IDS versus ADS Based on Vocal Timbre versus Background Noise

All "speech" (blue) bars are duplicated exactly from Figure 2 and appear again here for visual comparison. "Silence" (yellow) bars are derived from cropped segments containing no sounds except for ambient noise from recordings of English speakers and non-English speakers. Chance (dashed line) is 50%. N = 12. Classification performance is represented as mean percent correct and ±SEM across cross-validation folds (leave-one-subject-out). Figure S2 displays accuracy rates for classifying individual speakers based on speech versus background noise. *p < 0.05, ***p < 0.001.

silences alone for both English (Figure 4A; two-tailed, paired-samples t test, $t(11) = 2.92$, $p < 0.02$) and non-English (Figure 4B; $t(11) = 2.31$, $p < 0.05$) data. Furthermore, both of our cross-language analyses failed completely when we used silence alone (train on English silence, test on non-English silence: $t(11) = -1.38$, $p = 0.19$, see Figure 4C, yellow bar; train on non-English silence, test on English silence: $t(11) = -.04$, $p = 0.97$, see Figure 4D, yellow bar). And once again, cross-language classification of ADS versus IDS was significantly better when we trained on real English speech and tested on non-English speech than when we trained and tested on silence alone from those respective groups (Figure 4C; $t(11) = 6.46$, $p < 0.0001$) and was also better when we trained on non-English speech and tested on English speech than when we trained and tested on silence alone (Figure 4D; $t(11) = 2.59$, $p < 0.05$). Collectively, these results demonstrate that differences in background noise across ADS and IDS recordings cannot fully account for our ability to classify those two modes of speech.

DISCUSSION

We show for the first time that infant-directed speech is defined by robust shifts in overall timbre that help differentiate it from adult-directed speech as a distinct communicative mode across multiple languages. This spectral dimension of speakers' voices differs reliably between the two contexts, i.e., when a mother is speaking to her infant versus engaging in dialog with an adult. Our findings generalize across a broad set of languages, much as pitch characteristics of IDS manifest similarly in several languages [13]. This research emphasizes and isolates the significant role of timbre (a relatively high-level feature of sounds [14]) in communicative code switching, adding a novel dimension to the well-known adjustments that speakers use in IDS, such as higher pitch, more varied pitch, longer pauses, shorter utterances, and more repetition [3, 4].

Our study complements research showing that adult speakers acoustically exaggerate the formant frequencies of their speech when speaking to infants to maximize differences between vowels [12, 15]. Our results cannot be explained by differences in pitch (F0) or in the first two formants (F1 and F2) between ADS and IDS; even after regressing out these features over time from the MFCCs, significant timbre differences remain that allow for robust classification of these two modes of speech (see Results and Figure 3). This suggests that our timbre effects reflect a shift in the global shape of the spectrum beyond these individual frequency bands. Furthermore, the shifts we report generalize across a broader set of languages than have been tested in previous work. Even after the removal of pitch and formants, our ADS/IDS classification model transfers from English speech to speech sampled from a diverse set of nine other languages (Figures 3C and 3D, third bars).

Timbre enables us to discriminate, recognize, and enjoy a rich variety of sounds [5, 16], from friends' voices and animal calls to musical textures. A characteristic of the speech spectrum that depends on the resonant properties of the larynx, vocal tract, and articulators, vocal timbre varies widely across people (see Audio S1, S2, S3, and S4). Because MFCC has been shown to provide a strong model for perceptual timbre space as a whole [7], we focused here on this summary statistical measure of the spectral shape of a voice as a proxy for timbre. However, timbre is a complex property that requires the neural integration of multiple spectral and temporal features [17], most notably spectral centroid (the weighted average frequency of a signal's spectrum, which strongly relates to our MFCC measure and influences perceived brightness [18]), attack time, and temporal modulation (e.g., vibrato) [6]. Future work should explore how these dimensions of timbre might interact in IDS in order to support infants' learning of relevant units of speech.

Compared to most prosodic features of speech (e.g., pitch range, rhythm), which are more "horizontal" and unfold over multiple syllables and even sentences of data, the time-averaged summary statistic we have measured represents a more "vertical" dimension of sound. This spectral fingerprint is detectable and quantifiable with very little data, consistent with listeners' abilities to identify individual speakers even from single, monosyllabic words [19] and to estimate rich information from very brief clips of popular music [20]. These examples are too short to contain pitch variation but do include information about the relationships between harmonics that are linked to perceptual aspects of voice quality. Such cues include amplitude differences between harmonics, harmonics-to-noise ratio, and both jitter and shimmer (which relate to roughness or hoarseness [21]). Infants' abilities to classify [22] and remember [23] timbre suggest that it could be partly responsible for their early ability to recognize IDS [24] and their own caregivers' voices [25]. Both identification processes are likely to provide relevant input for further learning about the ambient language and the social environment. Because timbre contributes greatly to the rapid identification of sound sources and the grouping of auditory objects [5], it likely serves as an early and important cue to be associatively bound to other sensory features (e.g., a sibling with her voice, a dog with its bark). The developmental time course of this process invites future investigation.

The timbre shifts we report in IDS are likely part of a broadly adaptive mechanism that quickly draws infants' attention [26] to the statistical structure of their relevant auditory environment, starting very soon after birth [24], and helps them to segment words [27], learn the meanings of novel words [28], and segment speech into units [29]. IDS may serve as a vehicle for the expression of emotion [30], in part due to its "musical" characteristics [31, 32] and its interaction with a mother's emotional state [33]. One study [34] reported differences in several timbre-related acoustic features between infant-directed and adult-directed singing in English-speaking mothers, but it is unclear whether these differences are due to performance-related aspects of vocalization (i.e., having someone to interact with or not, which can affect speech behavior through non-verbal feedback [35]) or code switching between adult and child audiences. Because some timbre features have been shown to influence emotional ratings of speech prosody [36] and because affect is thought to mediate the learning benefits of IDS [33], future work might ask how the observed differences between IDS and ADS relate to mothers' tendencies to smile, gesture, or provide other emotional cues during learning.

Our findings have the potential to stimulate broad research on features of language use in a variety of communicative contexts. Although timbre's role in music has been widely studied [16], its importance for speech—and in particular, communicative signaling—is still quite poorly understood. However, timbre holds great promise for helping us to understand and quantify the frequent register shifts that are important for flexible communication across a wide variety of situations. For instance, performers often manipulate their timbre for emotional or comic effect, and listeners are sensitive to even mild affective changes signaled by these timbral shifts [37]. Future studies could expand existing literature on audience design [38–41] and code switching [42, 43] by exploring how speakers alter their timbre, or other vocal summary statistics, to flexibly meet the demands of a variety of audiences, such as friends, intimate partners, superiors, students, or political constituents.

Understanding how caregivers naturally alter their vocal timbre to accommodate children's unique communicative needs could have wide-ranging impact, from improving speech recognition software to improving education. For instance, our use of summary statistics could enable speech recognition algorithms to quickly and automatically identify infant-directed speech (and in the future, perhaps a diverse range of speech modes) from just a few seconds of data. This would support ongoing efforts to develop software that provides summary measures of natural speech in infants' and toddlers' daily lives through automatic vocalization analysis [44, 45]. Moreover, software designed to improve language or communication skills [46] could enhance children's engagement by adjusting the vocal timbre of virtual speakers or teaching agents to match the variation inherent in the voices of their own caregivers. Finally, this implementation of summary statistics could improve the efficiency of emerging sensory substitution technologies that cue acoustic properties of speech through other modalities [47, 48]. The statistics of timbre in different communicative modes have the potential to enrich our understanding of how infants tune in to important signals and people in their lives and to inform efforts to support children's language learning.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Equipment
  - General Procedure
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data Pre-processing
  - Timbre Analysis
  - Classification
  - Pitch and Formant Control Analyses
  - Background Noise Control Analysis
  - Classification of IDS versus ADS Cannot Be Explained by Differences between Read and Spontaneous Speech
  - Classification of Individual Speakers
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures, one table, and four audio files and can be found with this article online at https://doi.org/10.1016/j.cub.2017.08.074.

AUTHOR CONTRIBUTIONS

E.A.P. developed the study concept. E.A.P., M.C.I., and C.L.-W. contributed to the study design. E.A.P. collected data, and E.A.P and M.C.I. performed data analysis. E.A.P. drafted the manuscript, and M.C.I. and C.L.-W. provided critical revisions.

ACKNOWLEDGMENTS

This research was supported by a C.V. Starr Postdoctoral Fellowship from the Princeton Neuroscience Institute awarded to E.A.P. and a grant from the National Institute of Child Health and Human Development awarded to C.L.-W. (HD079779). The authors thank Julia Schorn and Eva Fourakis for assistance with data collection.

Received: May 11, 2017
Revised: July 25, 2017
Accepted: August 30, 2017
Published: October 12, 2017

REFERENCES

1. Ohala, J.J. (1983). Cross-language use of pitch: an ethological view. Phonetica *40*, 1–18.
2. Auer, P., Couper-Kuhlen, E., and Müller, F. (1999). Language in Time: The Rhythm and Tempo of Spoken Interaction (Oxford University Press on Demand).
3. Fernald, A. (1992). Human maternal vocalizations to infants as biologically relevant signals. In The Adapted Mind: Evolutionary Psychology and the Generation of Culture, J. Barkow, L. Cosmides, and J. Tooby, eds. (Oxford: Oxford University Press), pp. 391–428.
4. Fernald, A., and Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. Dev. Psychol. *20*, 104–113.
5. Bregman, A.S. (1994). Auditory Scene Analysis: The Perceptual Organization of Sound (Cambridge: MIT Press).
6. Elliott, T.M., Hamilton, L.S., and Theunissen, F.E. (2013). Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. J. Acoust. Soc. Am. *133*, 389–404.
7. Terasawa, H., Slaney, M., and Berger, J. (2005). Perceptual Distance in Timbre Space (Georgia Institute of Technology).
8. McKinney, M.F., and Breebaart, J. (2003). Features for audio and music classification. Proc. ISMIR *3*, 151–158.
9. Davis, S.B., and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE T. Acoust. Speech Signal Proc. *28*, 357–366.
10. Reynolds, D.A. (1994). Experimental evaluation of features for robust speaker identification. IEEE Trans. Speech Audio Process. *2*, 639–643.
11. Tiwari, V. (2010). MFCC and its applications in speaker recognition. Int. J. Emerg. Technol. *1*, 19–22.
12. 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 *277*, 684–686.
13. Grieser, D.L., and Kuhl, P.K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. Dev. Psychol. *24*, 14–20.
14. Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., and Possing, E.T. (2000). Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex *10*, 512–528.
15. Eaves, B.S., Feldman, N.H., Griffiths, T.L., and Shafto, P. (2016). Infant-directed speech is consistent with teaching. Psychol. Rev. *123*, 758–771.
16. McAdams, S., and Giordano, B.L. (2009). The perception of musical timbre. In Oxford Handbook of Music Psychology, S. Hallam, I. Cross, and M. Thaut, eds. (New York: Oxford University Press), pp. 72–80.
17. Patil, K., Pressnitzer, D., Shamma, S., and Elhilali, M. (2012). Music in our ears: the biological bases of musical timbre perception. PLoS Comput. Biol. *8*, e1002759.
18. Wessel, D.L. (1979). Timbre space as a musical control structure. Comput. Music J. *3*, 45–52.
19. Pollack, I., Pickett, J.M., and Sumby, W.H. (1954). On the identification of speakers by voice. J. Acoust. Soc. Am. *26*, 403–406.
20. Krumhansl, C.L. (2010). Plink: "Thin slices" of music. Music Percept. *27*, 337–354.
21. Kreiman, J., Gerratt, B.R., Precoda, K., and Berke, G.S. (1992). Individual differences in voice quality perception. J. Speech Hear. Res. *35*, 512–520.
22. Trehub, S.E., Endman, M.W., and Thorpe, L.A. (1990). Infants' perception of timbre: classification of complex tones by spectral structure. J. Exp. Child Psychol. *49*, 300–313.
23. Trainor, L.J., Wu, L., and Tsang, C.D. (2004). Long-term memory for music: infants remember tempo and timbre. Dev. Sci. *7*, 289–296.
24. Cooper, R.P., and Aslin, R.N. (1990). Preference for infant-directed speech in the first month after birth. Child Dev. *61*, 1584–1595.
25. DeCasper, A.J., and Fifer, W.P. (1980). Of human bonding: newborns prefer their mothers' voices. Science *208*, 1174–1176.
26. Kaplan, P.S., Goldstein, M.H., Huckeby, E.R., Owren, M.J., and Cooper, R.P. (1995). Dishabituation of visual attention by infant- versus adult-directed speech: Effects of frequency modulation and spectral composition. Infant Behav. Dev. *18*, 209–223.
27. Thiessen, E.D., Hill, E.A., and Saffran, J.R. (2005). Infant-directed speech facilitates word segmentation. Infancy *7*, 53–71.
28. Graf Estes, K., and Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. Infancy *18*, 797–824.

29. Kemler Nelson, D.G., Hirsh-Pasek, K., Jusczyk, P.W., and Cassidy, K.W. (1989). How the prosodic cues in motherese might assist language learning. J. Child Lang. *16*, 55–68.

30. Trainor, L.J., Austin, C.M., and Desjardins, R.N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? Psychol. Sci. *11*, 188–195.

31. Bergeson, T.R., and Trehub, S.E. (2002). Absolute pitch and tempo in mothers' songs to infants. Psychol. Sci. *13*, 72–75.

32. Bergeson, T.R., and Trehub, S.E. (2007). Signature tunes in mothers' speech to infants. Infant Behav. Dev. *30*, 648–654.

33. Kaplan, P.S., Bachorowski, J.A., and Zarlengo-Strouse, P. (1999). Child-directed speech produced by mothers with symptoms of depression fails to promote associative learning in 4-month-old infants. Child Dev. *70*, 560–570.

34. Trainor, L.J., Clark, E.D., Huntley, A., and Adams, B.A. (1997). The acoustic basis of preferences for infant-directed singing. Infant Behav. Dev. *20*, 383–396.

35. Blubaugh, J.A. (1969). Effects of positive and negative audience feedback on selected variables of speech behavior. Speech Monogr. *36*, 131–137.

36. Coutinho, E., and Dibben, N. (2013). Psychoacoustic cues to emotion in speech prosody and music. Cogn. Emotion *27*, 658–684.

37. Gobl, C., and Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood, and attitude. Speech Commun. *40*, 189–212.

38. Bell, A. (1984). Language style as audience design. Lang. Soc. *13*, 145–204.

39. Lam, T.Q., and Watson, D.G. (2010). Repetition is easy: why repeated referents have reduced prominence. Mem. Cognit. *38*, 1137–1146.

40. Rosa, E.C., Finch, K.H., Bergeson, M., and Arnold, J.E. (2015). The effects of addressee attention on prosodic prominence. Lang. Cogn. Neurosci. *30*, 48–56.

41. Jürgens, R., Grass, A., Drolet, M., and Fischer, J. (2015). Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. J. Nonverbal Behav. *39*, 195–214.

42. Koch, L.M., Gross, A.M., and Kolts, R. (2001). Attitudes toward Black English and code switching. J. Black Psychol. *27*, 29–42.

43. DeBose, C.E. (1992). Codeswitching: Black English and standard English in the African-American linguistic repertoire. J. Multiling. Multicul. Dev. *13*, 157–167.

44. Xu, D., Yapanel, U., and Gray, S. (2009). Reliability of the LENA language environment analysis system in young children's natural home environment. www.lenafoundation.org/wp-content/uploads/2014/10/LTR-05-2_Reliability.pdf

45. Ford, M., Baer, C.T., Xu, D., Yapanel, U., and Gray, S. (2009). The LENA language environment analysis system: Audio specifications of the DLP-0121. https://pdfs.semanticscholar.org/44d1/08871b090c846d40fb1c096cdd279a627c2c.pdf

46. Bosseler, A., and Massaro, D.W. (2003). Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. J. Autism Dev. Disord. *33*, 653–672.

47. Saunders, F.A. (1983). Information transmission across the skin: high-resolution tactile sensory aids for the deaf and the blind. Int. J. Neurosci. *19*, 21–28.

48. Massaro, D.W., Carreira-Perpinan, M.A., Merrill, D.J., Sterling, C., Bigler, S., Piazza, E., and Perlman, M. (2008). iGlasses: an automatic wearable speech supplement in face-to-face communication and classroom situations. Proc. ICMI *10*, 197–198.

49. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2006). The HTK Book (HTK Version 3.4.1) (Cambridge University).

50. Chang, C.C., and Lin, C.J. (2011). LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. *2*, 27.

51. Boersma, P., and Weenink, D. (2009). Praat: doing phonetics by computer, Version 6.0.14 (Princeton University).

52. Koreman, J., Andreeva, B., and Strik, H. (1999). Acoustic parameters versus phonetic features in ASR. Proc. Int. Congr. Phonet. Sci. *99*, 719–722.

53. Van den Oord, A., Dieleman, S., and Schrauwen, B. (2013). Deep content-based music recommendation. In Adv. Neur. In. Proc. Sys. 2643–2651.

54. Howell, P., and Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spontaneous speech material. Speech Commun. *10*, 163–169.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
MATLAB R2016A	MathWorks	RRID: SCR_001622
HTK MFCC toolbox	[9, 49]	http://www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab
LibSVM	[50]	https://www.csie.ntu.edu.tw/~cjlin/libsvm/
Praat	[51]	http://www.fon.hum.uva.nl/praat/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Elise Piazza (elise.piazza@gmail.com).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Twenty-four mother-infant dyads participated in three naturalistic activities (see General Procedure), the order of which was counterbalanced across participants. Infants were 7-12 months old. Informed consent was obtained from all participating mothers, and approval of the study was obtained from the Princeton University Institutional Review Board. Participants were given no information about the experimental hypotheses; that is, they were told that we were broadly interested in ‘‘how mothers interact with their infants’’ and were not aware that we were measuring differences between the acoustic properties of their speech across the conditions. We chose to test only mothers to keep overall pitch range fairly consistent across participants but would expect these results to generalize to fathers as well, which could be explored in future studies.

Twelve of the mother-infant dyads were English speakers. To investigate the possibility that timbral differences between ADS and IDS generalize across languages, we recorded a second group of 12 mothers who speak to their infants using a language other than English at least 50% of the time. We included a wide variety of languages: Spanish (N = 1), Russian (N = 1), Polish (N = 1), German (N = 2), Hungarian (N = 1), French (N = 1), Hebrew (N = 2), Mandarin (N = 2), Cantonese (N = 1). All families were recruited from the central New Jersey area.

METHOD DETAILS

Equipment

Speech data were recorded continuously using an Apple iPhone and a Miracle Sound Deluxe Lavalier lapel microphone attached to each mother’s shirt. Due to microphone failure, three participants were recorded with a back-up Blue Snowball USB microphone; recording quality did not differ between microphones.

General Procedure

In the adult-directed speech (ADS) condition, mothers were interviewed by an adult experimenter about the child’s typical daily routine, feeding and sleeping habits, personality, and amount of time spent with various adults and children in the child’s life. In the two infant-directed speech (IDS) conditions, mothers were instructed to interact freely with their infants, as they naturally would at home, by playing with a constrained set of animal toys and reading from a set of age-appropriate board books, respectively. Each condition lasted approximately 5 min. See Table S1 for example utterances from both conditions.

All procedures were identical for the English-speaking and non-English-speaking mothers, except that non-English-speaking mothers were asked to speak only in their non-English language during all experimental conditions (adult interview, reading, and play). In the interview condition, the experimenter asked the questions in English, but the mother was asked to respond in her non-English language, and she was told that a native speaker of her non-English language would later take notes from the recordings. We chose this method (instead of asking the participants to respond to a series of written questions in the appropriate language) to approximate a naturalistic interaction (via gestures and eye contact) as closely as possible.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data Pre-processing

Using Adobe Audition, we extracted 20, two-second phrases from each condition (ADS, IDS) for each mother. Phrases were chosen to represent a wide range of semantic and syntactic content; see Table S1 for example phrases. Data were manually inspected to ensure that they did not include any non-mother vocalizations or other extraneous sounds (e.g., blocks being thrown). After excluding these irrelevant sounds, there were typically only 20-30 utterances to choose from, and we simply chose the first 20 from each mother.

Timbre Analysis

To quantify each mother's timbre signature, we used Mel-frequency cepstral coefficients (MFCC), a measure of timbre used for automatic recognition of individual speakers [10, 11], phonemes [52] and words [9], and musical genres [8, 53]. The MFCC is a feature set that succinctly describes the shape of the short-term speech spectrum using a small vector of weights. The weights represent the coefficients of the discrete cosine transform of the Mel spectrum, which is designed to realistically approximate the human auditory system's response. In our analysis, the MFCC serves as a time-averaged summary statistic that describes the global signature of a person's voice over time. For each phrase, we first computed the MFCC in each 25-ms time window to yield a coefficient x time matrix. MFCCs were computed using 25-ms overlapping windows, each 10 ms apart. Finally, we computed a single, time-averaged vector, consisting of 12 MFCC coefficients, across the entire duration of the phrase (Figure 1). Figure S1 shows the average ADS and IDS vectors across all 20 phrases for each English-speaking participant. The MFCC features were extracted according to [49] and [9], as implemented in the HTK MFCC analysis package in MATLAB R2016A, using the following default settings: 25-ms overlapping windows, 10-ms shift between windows, pre-emphasis coefficient = 0.97, frequency range = 300 to 3700 Hz, 20 filterbank channels, 13 cepstral coefficients, cepstral sine lifter parameter = 22, hamming window.

Classification

Prior work suggests that timbre as a feature of natural speech is consistent within individuals and distinct between individuals [19], i.e., that timbral signatures possess enough information to distinguish individual mothers from one another. As an initial validation of our method, we first confirmed that support-vector machine (SVM) classification is sensitive enough to replicate previous work distinguishing individual mothers [10, 11] by performing the classification on these MFCC vectors across subjects (see Figure S2). Then, to test our primary question of interest, we performed a similar SVM classification on these vectors to distinguish IDS from ADS.

To this end, we used a support vector machine classifier with radial basis function kernel (SVM-RBF) (LibSVM implementation [50],) to predict whether utterances belong to the IDS or ADS communicative modes based on MFCC features extracted from natural speech. We employed a standard leave-one-subject-out (LOSO) 12-fold cross-validation, where for each cross-validation fold we trained our classifier to distinguish between IDS and ADS using the data from 11 subjects and then tested how well it generalized to the left-out (twelfth) subject. We note that this is a more stringent test for our classifier compared to using training data from each subject, because in our case the classifier is oblivious to idiosyncrasies of the left-out subject's speech when building a model to discriminate between IDS and ADS.

Additionally, within each LOSO cross-validation fold, we also performed a log-space grid search in the (2^{-14}, 2^{14}) range on a randomly selected held-out subset of one quarter of the training data (25% x 91.7% = 22.9% of the original data) to select the optimal classifier parameters (C, Gamma). After selecting the optimal cross-validated parameters for that particular fold, we re-trained our classifier on the entire fold's training set (11 subjects) and tested how well it generalized to the left-out (twelfth) subject from the original data. We repeated this procedure 12 times, iterating over each individual subject for testing (e.g., the first subject is held out for fold #1 testing, the second subject is held out for fold #2 testing, etc.).

We performed the same analysis described above (classify IDS versus ADS) on a second group of mothers who spoke a language other than English during the experimental session. In addition, to test the generalizability of the timbre transformation between ADS and IDS across languages, we performed cross-language classification. Specifically, we first trained the classifier to distinguish between IDS and ADS using data from the English participants only and tested it on data from the non-English cohort. Finally, we performed the reverse analysis, where the classifier was trained on non-English data and was used to predict IDS versus ADS in the English participant cohort. SVM classification results (mean percent correct and ± SEM across cross-validation folds) are shown in Figure 2.

Pitch and Formant Control Analyses

We performed a control analysis to rule out the possibility that our ability to distinguish IDS from ADS was based on differences in pitch that were somehow recoverable in the MFCC feature set. For every utterance in our dataset, we used Praat [51] to extract a vector corresponding to the entire F0 contour. For time points in which no F0 value was estimated (i.e., non-pitched speech sounds), we removed these samples from the F0 vector and also removed the corresponding time bins from the MFCC matrix (coefficients x times; see Timbre Analysis). This temporal alignment of the MFCC matrices and F0 vectors allowed us to regress out the latter from the former. Classification performance for these time-restricted MFCC matrices, before regressing anything out, is shown in Figure 3 (first bars). Next, we regressed out F0 (over time) from each of the 12 MFCC time vectors before computing the single time-averaged vector of MFCC coefficients for each utterance and performing SVM classification between ADS and IDS based on these residual time-averaged vectors (Figure 3, second bars).

We performed a very similar analysis to additionally control for the impact of the first two formants (F1 and F2) on our classification of IDS and ADS. Specifically, for each utterance, we used Praat [51] to extract two vectors corresponding to F1 and F2; these vectors represented the same time points as the F0 vector described above to ensure temporal alignment for the purposes of regression. We then regressed out F0, F1, and F2 (over time) from each of the 12 MFCC time vectors before computing the single time-averaged vector of MFCC coefficients and performing classification between ADS and IDS based on these residual time-averaged vectors (Figure 3, third bars).

Background Noise Control Analysis

We performed a control analysis to rule out the possibility that our ability to distinguish between IDS and ADS was due to differences between the noise properties of the microphone or room across different conditions (Figure 4). Here, instead of utterances, we extracted 20 segments of silence (which included only ambient, static background noise and no vocalizations, breathing, or other dynamic sounds) from the ADS and IDS recordings of each participant, of comparable length to the original speech utterances (1-2 s). We then performed identical analyses to those described above for speech data (i.e., analyzing the MFCC of each segment of silence, performing cross-validated SVM classification to discriminate IDS from ADS recordings).

Classification of IDS versus ADS Cannot Be Explained by Differences between Read and Spontaneous Speech

We conducted a control analysis to ensure that known prosodic differences between read and spontaneous speech [54] could not account for our ability to distinguish IDS from ADS data. Specifically, for each of the 12 English-speaking mothers, we replaced all IDS utterances that corresponded to the "book" (reading) condition with new utterances from the same mother's recording that corresponded to the "play" condition only. Thus, all 20 utterances from both IDS and ADS now represented only spontaneous speech. Resulting classification remained significantly above chance (two-tailed, one-sample t test, $t(11) = 9.81$, $p < 0.0001$), indicating that potential differences between spontaneous and read speech could not account for our results.

Classification of Individual Speakers

To confirm that our method is sufficiently sensitive to distinguish between different participants, as in previous research [10, 11], we used a similar classification technique as the one used to compare IDS to ADS. More specifically, we employed a standard "leave-two-utterances-out" 10-fold cross-validation procedure for testing, where for each fold we left out 10% of the utterances from each condition and each subject (e.g., two utterances each from IDS and ADS per subject) and trained the classifier on the remaining 90% of the data, before testing on the left-out 10%. Within each fold, we also performed a log-space grid search in the (2^{-14}, 2^{14}) range on a held-out subset of one quarter of the training data (25% x 90% = 22.5% of the original data) to select the optimal classifier parameters (C, Gamma). After selecting the optimal cross-validated parameters for that particular fold, we re-trained our classifier on the entire training set (90% of original data) and tested how well it generalized to the left-out 10% of the original data. We repeated this procedure 10 times, iterating over non-overlapping subsets of held-out data for testing (e.g., the first two utterances from IDS and ADS are held out for fold #1 testing, the next two utterances are held out for fold #2 testing, etc.).

Using this procedure, we were able to reliably distinguish between individual mothers significantly above chance based on MFCCs from English speech data (Figure S2A, blue bar; two-tailed, one-sample t test, $t(11) = 23.26$, $p < 0.0001$) and from non-English speech data (Figure S2B, blue bar; $t(11) = 22.22$, $p < 0.0001$).

We also performed another control analysis based on background noise (similar to the one above) to rule out the possibility that our ability to distinguish between different individuals was due to differences between the noise properties of the microphone or room across different days. We found that we could classify individual mothers above chance in both the English-speaking group (Figure S2A, yellow bar; $t(11) = 19.42$, $p < 0.0001$) and the non-English-speaking group (Figure S2B, yellow bar; $t(11) = 26.78$, $p < 0.0001$) based on silences in the recordings alone. This is not entirely surprising because although we maintained a consistent distance (approximately 12 inches) from the mouth to microphone across mothers, the background noise conditions of the room may have changed slightly beyond our control across days (e.g., due to differences in the settings of the heating unit). Importantly, however, discrimination of individual speakers was significantly better for real speech than silence segments, for both English (Figure S2A; two-tailed, paired samples t test, $t(11) = 4.52$, $p < 0.001$) and non-English data (Figure S2B; $t(11) = 3.45$, $p < 0.01$).

DATA AND SOFTWARE AVAILABILITY

Interested readers are encouraged to contact the Lead Contact for the availability of data. The MATLAB-based MFCC routines can be found at: http://www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab.

Current Biology, Volume *27*

Supplemental Information

Mothers Consistently Alter Their Unique Vocal Fingerprints When Communicating with Infants

Elise A. Piazza, Marius Cătălin Iordan, and Casey Lew-Williams

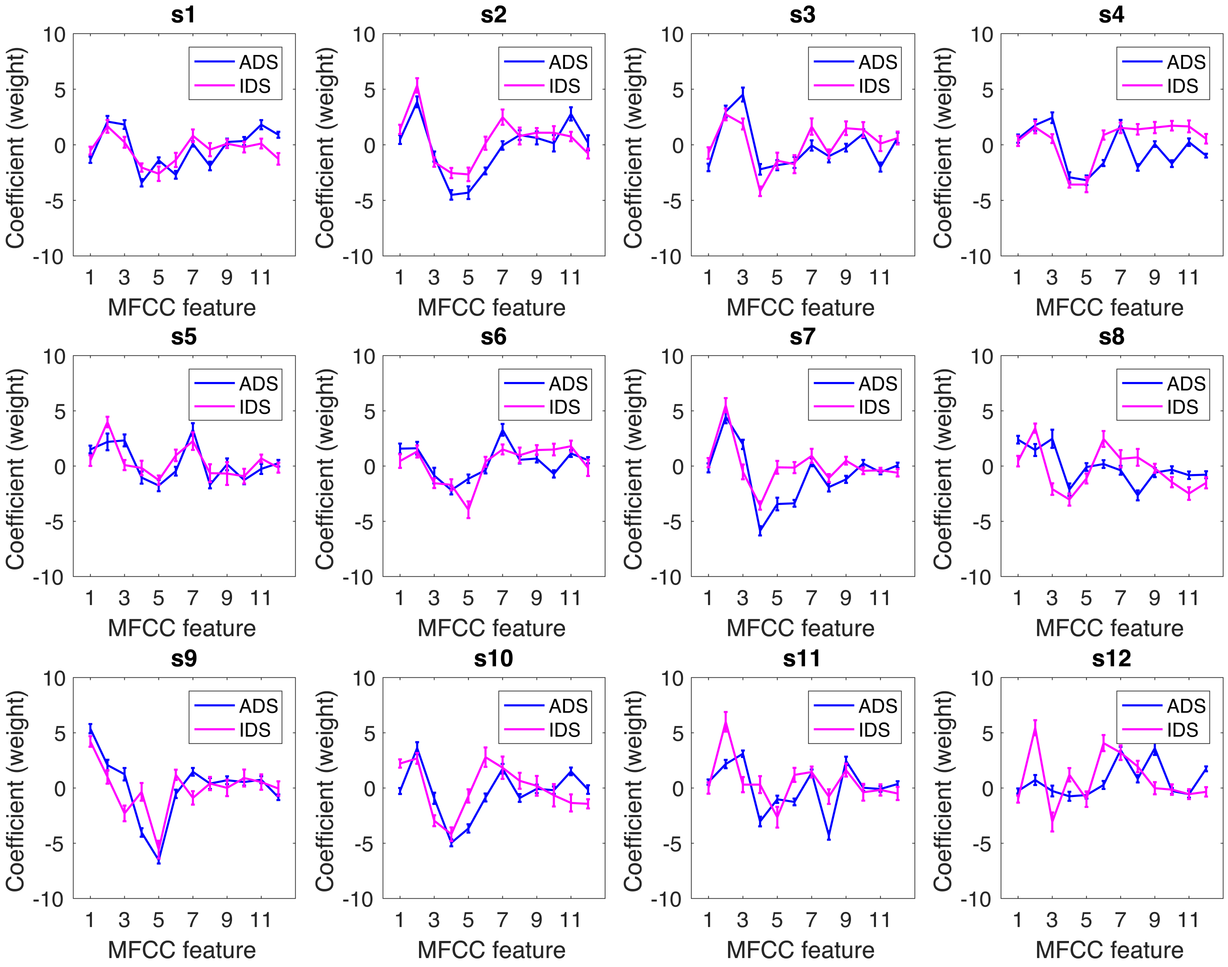

Figure S1. Average MFCC feature vectors across all utterances for each English-speaking participant; related to Figure 1. Each plot represents data from one participant's adult-directed and infant-directed speech. N = 12. Error bars represent ± SEM across 20 utterances.

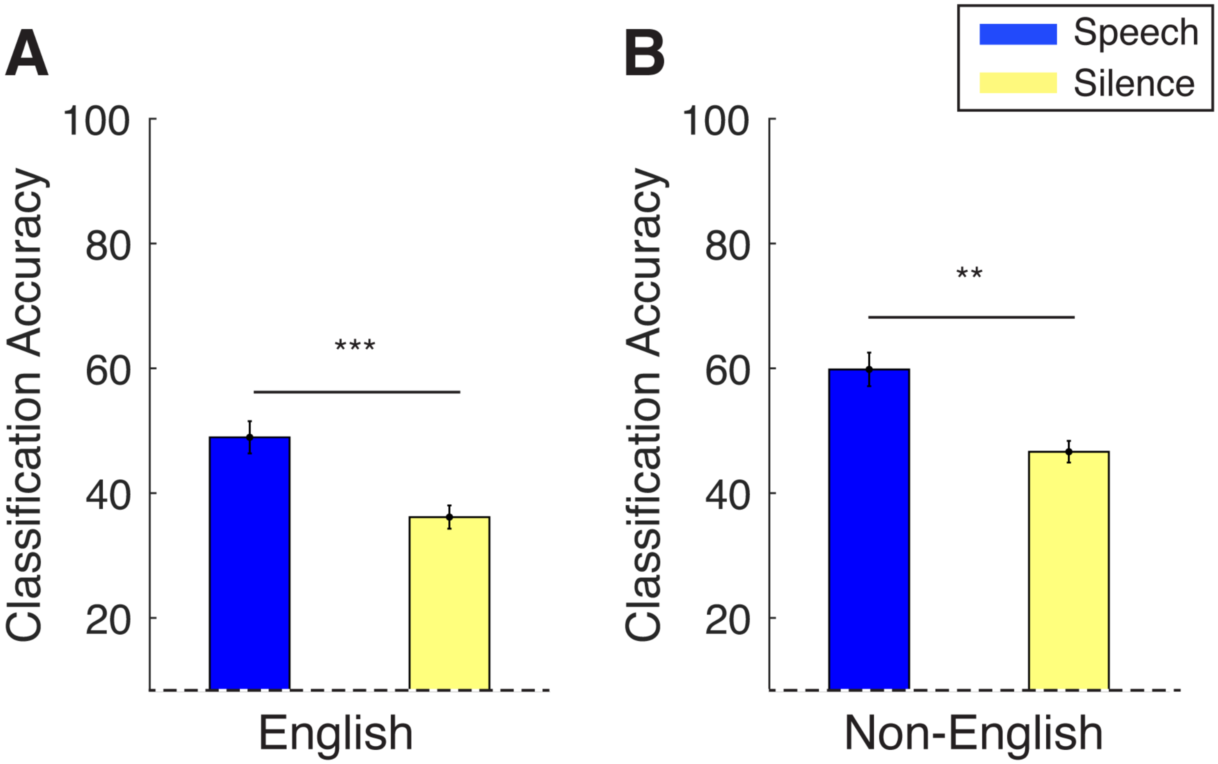

Figure S2. Accuracy rates for classifying individual subjects based on vocal timbre vs. background noise; related to Figure 4 and STAR Methods. “Silence” (yellow) bars are derived from cropped segments containing no sounds except for ambient recording noise. Chance (dashed line) is 1/N (8.33%). N = 12. Classification performance is represented as mean percent correct and ± SEM across cross-validation folds. ***p* < .01, ****p* < .001.

Adult-directed utterances	Infant-directed utterances
They had to sleep in their crib	*Say ‘night, night, puppy’*
Oh my goodness, a typical day	*No, we don’t want to eat the birdie*
Yeah, but she seems to enjoy reading	*I know—the light is still up there*
Four-thirty to six-thirty	*Flap, flap, flap, over the field*
He’s looking for food and nothing else	*Three little bears, sitting on chairs*
We’ll play for a little while	*Look—he’s scared behind the tree*
He does like to look at the pictures	*They’re in the water*
Yeah, I mean, they’re kind of regular	*The big thing just said ‘snort’*
He loves his little walker toy	*One, two, three*
He’s drinking by himself now	*Put the animals in there*
Take a bath and then get him dressed	*Bunny…where’s the bunny?*
He’s usually pretty fine; he loves to be out	*Let’s not eat the kitty cat*
So, we don’t really have a routine like other families	*Where’s that doggy at?*
I mean, we read every night	*Let’s dump it all out*
My husband’s around a lot	*Your sister’s first word—did you know that?*
Gone for those twenty-four hours	*The box is exciting, isn’t it?*
So, I work Saturdays, so my husband has off	*Hit the cat and the cow together?*
But previous to that, he just kind of brings, like, toys	*Look at this golden retriever!*
We usually Skype with my parents	*Is the doggy yummy?*
And then we go downstairs and eat our breakfast	*A mother bird sat on her eggs*
We have breakfast and we usually talk to my mum	*They jumped and jumped and jumped*
A parents’ group that the hospital runs	*We haven’t visited goats in a while*
Today it was for me; it was a crochet class	*In comparison to these other-sized animals*
She sees her dad, like, for ten minutes in the morning	*It’s a little house for them*
He was so into the movement	*Do we have deers behind our house?*
He’s with me all the time	*There are many things that I like about being me*
It took him nine months	*That was ambitious!*
Yeah, we had some bad times	*But who loves to boogie?*
And that’s his favorite—he’s got shelves	*The kitten just looked and looked*
Like, not to whack each other, and things like that, too	*And a pair of mittens*

Table S1. Sample utterances of English speech data used in analyses; related to STAR Methods. Each mother contributed 2-3 utterances.