The emotional component of Infant Directed-Speech: a cross-cultural study using machine learning

Brief title: IDS across languages

Erika Parlato-Oliveira[1], Mohamed Chetouani[2], Jean-Maximilien Cadic[2], Sylvie Viaux[2,3], Zeineb Ghattassi[2,3], Jean Xavier[2,3], Lisa Ouss[4], Ruth Feldman[5], Filippo Muratori[6], David Cohen[2,3], Catherine Saint-Georges[2,3]

[1] Federal University of Minas Gerais, Faculty of Medicine, Belo Horizonte, Brazil; Université Paris Diderot, 75013 Paris, France
[2] Institut des Systèmes Intelligents et de Robotiques, CNRS UMR 7222, Sorbonne Université, 75005 Paris, France
[3] Département de Psychiatrie de l'Enfant et de l'Adolescent, AP-HP.Sorbonne Université, Hôpital Pitié-Salpêtrière, 75013 Paris, France
[4] Service de Psychiatrie de l'Enfant, AP-HP, Hôpital Necker, 75015 Paris, France
[5] Center for Developmental Social Neuroscience and Baruch Ivcher School of Psychology, Interdisciplinary Center Herzliya, Herzliya, Israel.
[6] IRCCS Scientific Institute Stella Maris, University of Pisa, Calambrone (PI), Italy

*Corresponding Author (david.cohen@aphp.fr)
Département de Psychiatrie de l'Enfant et de l'Adolescent, AP-HP, Hôpital Pitié-Salpêtrière, 47-83, Boulevard de l'Hôpital, 75651 Paris, Cedex 13, France

RESUMÉ
Contexte : Le mamanais (ou infant-directed speech) entre dans une spirale interactive qui joue un rôle important dans le développement des compétences cognitives et sociales du bébé. L'usage du mamanais est universel et comprend des éléments linguistiques et émotionnels. Cependant, la question de la similarité ou non des composantes émotionnelles au niveau acoustique n'a jamais été explorée automatiquement.
Matériels et méthodes : Nous avons mené une étude transculturelle utilisant des techniques automatisées de traitement du signal social afin de comparer le mamanais dans différentes langues. Notre corpus de paroles est composé de vocalisations de parents enregistrées pendant des interactions avec leurs bébés âgés de 4 à 18 mois. Il inclut 6 bases de données en cinq langues: anglais, français, hébreu (deux bases de données : mères / pères), italien et portugais brésilien. Nous avons utilisé un classifieur automatique qui exploite les caractéristiques acoustiques de la parole et des méthodes de machine learning (apprentissage automatique par machines à vecteurs de support, SVM) afin de distinguer le mamanais-émotionnel du mamanais-non émotionnel.
Résultats : La classification automatique du mamanais émotionnel a été possible pour toutes les langues et tous les locuteurs (père et mère). La condition mono-langage (classifieur entrainé et testé dans la même langue) a produit des résultats de classification modérés à excellents, tous significativement supérieurs au hasard ($p < 1x10^{-10}$). Plus intéressant encore, la condition croisée (classifieur entrainé dans une langue et testé dans une autre langue) a produit des résultats de classification significativement supérieurs au hasard ($p < 1x10^{-10}$).
Conclusion : La classification automatique des composants émotionnels et non émotionnels du mamanais est possible sur la base des caractéristiques acoustiques indépendamment de la langue. Les résultats trouvés en condition croisée supportent l'hypothèse selon laquelle la composante émotionnelle repose sur des caractéristiques acoustiques similaires quelle que soit la langue.
Mots-clés: mamanais; interaction mère-bébé; transculturel; machine learning; traitement du signal social

ABSTRACT
Backgrounds: Infant-directed speech (IDS) is part of an interactive loop that plays an important role in infants' cognitive and social development. The use of IDS is universal and is composed of linguistic and emotional components. However, whether the emotional component has similar acoustics characteristics has not been studied automatically.
Methods: We performed a cross-cultural study using automatic social signal processing techniques (SSP) to compare IDS across languages. Our speech corpus consisted of audio-recorded vocalizations from parents during interactions with their infant between the ages of 4 and 18 months. It included 6 databases of five languages: English, French, Hebrew (two databases: mothers/fathers), Italian, and Brazilian Portuguese. We used an automatic classifier that exploits the acoustic characteristics of speech and machine learning methods (Support Vector Machines, SVM) to distinguish emotional IDS and non-emotional IDS.
Results: Automated classification of emotional IDS was possible for all languages and speakers (father and mother). The uni-language condition (classifier trained and tested in the same language) produced moderate to excellent classification results, all of which were significantly different from chance ($p<1x10^{-10}$). More interestingly, the cross-over condition (IDS classifier trained in one language and tested in another language) produced classification results that were all significantly different from chance ($p<1x10^{-10}$).
Conclusion: The automated classification of emotional and non-emotional components of IDS is possible based on the acoustic characteristics regardless of the language. The results found in the cross-over condition support the hypothesis that the emotional component shares similar acoustic characteristics across languages.
Keywords: motherese; mother-child interaction; cross-cultural; machine learning; social signal processing

Introduction

The spontaneous way in which mothers, fathers, and caregivers speak with infants and young children has been studied extensively across a number of interactive situations and contexts [1,2]. It is known as infant-directed speech (IDS) or “baby-talk” or motherese. Since Ferguson’s seminal study (1964, [3]), IDS was first studied by researchers who were interested in understanding language acquisition. It was found some general linguistic characteristics (e.g., shorter, linguistically simpler, redundant utterances) and prosodic characteristics (e.g., longer pauses, a slower tempo, more prosodic repetitions, and a higher mean f0) in a variety of languages [4]. Also, IDS is not restricted to mothers or women, and can be found in fathers as well. The phonological, lexical and syntactic properties of IDS contribute to infants’ language acquisition and comprehension [4]. IDS also has affective properties and contributes to the regulation of caregiver-infant interactions [5]. Both interaction partners can influence IDS. Caregivers’ characteristics (e.g., cultural, psychological and physiological ones) can influence IDS. Similarly, infant characteristics (e.g., reactivity and interactive feedback, age, and developmental condition) is associated with IDS significant variation [5,6]. Additionally, during early interaction, infants’ affect, attention and language learning are triggered by IDS. During interactions, some authors pointed that IDS supra-linguistic characteristics (e.g. prosody) reflect emotional charges of the caregiver and align with infants’ preferences [6]. In addition, mother-infant contingency and synchrony – two major social signals in both animals and humans – are crucial for IDS production and maintenance. Thus, IDS plays an important role in infants’ cognitive and social development, and is part of an interactive loop between caregivers and infants [6]. Recently Spinelli et al. [7] distinguishing prosodic features of IDS (F0 mean, F0 variability, F0 contour) conducted a meta-analysis to explore IDS influence on several infant outcomes and how some moderators could be involved. They showed an overall association between IDS prosody and better infant outcomes. This association was confirmed for attentional, prelinguistic and linguistic outcomes with a greater effect on pre-linguistic. Moderators were age (with a greater association for infant less than 9 months) and F0 features (with a greater effect with F0 contour). However, other acoustic parameters that define prosody were only occasionally explored in IDS [8] to be included in a meta-analysis.

From a broader communication perspective, IDS might be part of a more general phenomenon of adaptation to a partner during communication. First, cross-linguistic studies tented to support its universality [4]: across different languages, the same types of contours convey the same types of meanings,

which include arousing/soothing, turn-opening/turn-closing, approving/disapproving, and didactic modeling [9]. Second, mothers who are communicating in sign languages show differences in their signing to infants compared to adults, including simpler forms, a slower tempo, and more exaggerated movements [4]. Third, the use of IDS by humans has been compared to the “caregiver call” (which is almost exclusively infant-directed) in squirrel monkeys, in which the variability of several acoustic features, most notably pitch range and contour, is associated with particular contexts of infant care, such as nursing or retrieval [10]. Similarly, tamarins are calmed by music with the “acoustical characteristics of tamarin affiliation vocalizations” [11]. In a comparison of the mother-infant gestural and vocal interactions of chimpanzees and humans, Falk suggested that pre-linguistic vocal substrates for IDS evolved as females gave birth to relatively undeveloped neonates and adopted new strategies that entailed maternal quieting, reassuring, and controlling of the behaviors of physically removed infants (who were unable to cling to their mother’s body) [12]. The characteristic of vocal melodies of human mothers' speech to infants might be biologically relevant signals that have been shaped by natural selection [13], a finding that is integrated in the more general human and non-human communication field.

Several authors [12,14] recognized that IDS is universal but insist on the composition and frequency of its specific features that show variations across cultures [14]. A key example comes from Quiché-Mayan IDS, which does not have the typical pitch raising that is found in most IDS [15]. Therefore, IDS can be divided in non-emotional (linguistic) and emotional components. However, whether the emotional component has similar acoustics characteristics across languages and culture has not been studied automatically. The differential values of IDS emotional component vs. its non-emotional one has been studied by Matsuda et al. [16]. They proposed an original setting using neuroimaging and different stimuli with variation of both the supra-linguistic/emotional and the linguistic components of IDS. They showed that mothers of toddlers at the two-word stage who use both non-emotional (linguistic) and emotional components of IDS exhibited a significant interaction between the linguistic and emotional components of IDS mediated by the right caudate nucleus.

A way to contribute to this open debate is to perform cross-cultural studies using automatic social signal processing (SSP) techniques. SSP employs computational and cross-disciplinary approaches to investigate social interaction [17]. These techniques have been successfully applied to identify (i) speech and language based markers of social interaction in autism [18,19], (ii)

nonverbal cues of depression [20] and (iii) interpersonal cues of parent-infant interaction [21,22,23].

In this paper, we propose the use of an automatic classifier that has been developed to categorize speech segments of caregivers into emotional IDS sequences and non-emotional IDS sequences [8]. This method allowed us to study in home movies, caregivers' adaptation when interacting with an infant who will later develop autism [5]. Here, we aimed to study several databases of IDS with a machine learning algorithm based on acoustic components both in uni-language condition (IDS classifier trains and tests in the same language) and a cross-over language condition (IDS classifier trains in one language and tests in another language). We argue that a machine learning approach will help us to discover acoustic similarities among the databases, which in turn can inform us about the emotional characteristics of IDS across languages. In terms of hypothesis, we expect machine learning metrics, particularly the accuracy, to be above chance levels when classifying emotional characteristics of across languages [6]. We hypothesize similar results despite (1) intrinsic language characteristics (e.g. English vs. Hebrew) [8,24-25], (2) speaker traits (e.g. mother vs. father) [5,16], (3) cultural traits that can partially be expressed through language characteristics [27-29], and (4) audio recording conditions (e.g. home movie vs. audio lab recording) [8]. As working hypotheses, we expect that testing the same language as the one used during training will result in the best emotional IDS classifier metrics (hypothesis 1). Regarding points 1 and 3, we expect that testing another language than the one used during training should yield lower metrics in general but still above the chance level (hypothesis 2). This cross-over conditions will allow us to evaluate the generalization capabilities of the emotional IDS classifier effectively in terms of acoustic properties across languages. Similarly, regarding point 2, we expect that changing speakers during training and testing or mixing speakers in a given database should still yield metrics above the chance level (hypothesis 3).

Methods

Databases employed in the current study (Table1)

In the current study, we used 6 databases corresponding to five different languages: English, French, Hebrew, Italian, and Brazilian Portuguese. For Hebrew, we had two sets of data coming from two different experiments: one from mothers (Hebrew M) and one from fathers (Hebrew F). The Hebrew

maternal database was provided by 53 mothers who were playing freely with their six-month-old infant who was sitting in a relaxed position during a still-face experiment [30]. The 32 Hebrew fathers were participating in a cross-over trial testing the influence of oxytocin vs. placebo on father-infant interaction [31,32]. As in the Hebrew M database, the fathers were freely playing with their four-to seven-month-old infant during a still-face experiment. In this study, we only used the first free play period before still-face as it appears the more spontaneous period. In the two experiments, audio registration was of good quality as performed in an experimental room with a specific microphone per partner. The Italian data consisted of caregivers' (70% mothers, 30% fathers) speech to their 3- to 18-month child. They were extracted from the Pisa database of family home movies that were spontaneously made at home by the parents of children who later had typical development or developed autism. All the home movies were made before any medical diagnosis [5]. Audio registration was of poor quality as performed through microphones included on the video recorders used by families. The English database consisted of mothers' IDS that was collected from the Baby-Ears dataset. The database is publicly available. The data were collected in an experiment with 6 mothers who were speaking to their 10- to 18-month-old child while interacting with toys and preventing them from approaching some "dangerous" items [33]. The French database consisted of 5 mothers who were speaking to their 6- to 12-month-old child in the experimental context of the PILE project in which the mothers were asked to play and interact spontaneously with their child who was sitting in a baby-relax [34]. Sequences of interaction lasted 3- to 5-min and audio recording was of good quality as using a microphone in an experimental room. The Portuguese database was produced by 13 Brazilian mothers who were playing with their infant in a similar protocol. The Brazilian dyads were recruited between October 2012 and February 2013 in the Department of Psycholinguistics at the Universitade Federal de Mina Gerais (Belo Horizonte, Brazil) in an ongoing protocol to investigate mother-infant communication during the first year. Early interactions in free play were recorded at the local university hospital. All the parents gave their written informed consent to participate in the study, including the video recording (local ethics committee number: CAAE-0357.0.203.0000-11).

Except for the Baby-Ears dataset that is made publicly available with segmented sequences of mother's IDS, the following pre-processing was needed to perform the SSP studies. For each database, we manually segmented the parents' vocalizations from the video segments with ELAN [31]. Vocal segments of poor audio quality were excluded. The speech segments were

typically between 0.5 seconds and 4 seconds in length. In total, we had 2322 segments for English-speaking mothers, 3490 segments for Hebrew-speaking mothers, 1994 segments for Hebrew-speaking fathers, 35504 segments for Italian-speaking mothers and fathers, 5734 segments for Portuguese Brazilian-speaking mothers, and 3622 segments for French-speaking mothers.

The verbal interactions of the caregivers were carefully annotated by two blind psycholinguists into two categories: emotional IDS or non-emotional IDS. For each language, one psycholinguist had to be a native speaker. Both were blind to the purpose of the current study. We developed this approach in Italian first (see [8]) and showed that the more efficient detection of emotional IDS requires the selection of prototypic segments for training. This was achieved by excluding all the caregivers' vocalizations that received a discrepant annotation by the two raters. Finally, due to the different sizes of database we used, the number of selected sequences differed greatly from language to language (see table 1). Consequently for classification needs, we randomly selected the same number of sequences for each category (emotional IDS vs. non-emotional IDS) and in each language so that the chance level was equal to 0.5.

Automatic detection of the IDS emotional component

The automatic recognition of human emotions from speech signals has been an active field of research over the last decade. State-of-the-art methods involve (i) extracting features from speech signals, (ii) learning to predict training labels, and (iii) evaluating the system using a test database. Various features can be extracted from speech signals. Here, we used the INTERSPEECH 2013 Computational Paralinguistics Evaluation (ComParE) feature set [35]. This set includes 4 energy, 41 spectral, 14 cepstral (MFCC) and 6 voice-related low-level descriptors (LLDs). They are summarized in table 2.

A variety of functions were applied to the LLDs and delta LLDs to summarize the evolution of the contours over time. We used OpenSMILE (Speech and Music Interpretation by Large-space Extraction [36]) to extract the previously described 6,373 features. For the prediction step, different machine learning algorithms can be used. In this paper, a linear support vector machine (SVM) classifier is generated in a one-to-one strategy using a 10-fold cross-validation approach. To evaluate the generalization capabilities of our IDS emotional component classifier, we developed several learning and testing strategies (Figure 1):

• One-to-one condition: This is a within-corpus testing approach that requires

that the classifier is both machine-learned and tested using speech segments from one of the corpuses.
• Cross-over-condition: We exploit an off-corpus testing approach in which the classifier is first machine-learned using one corpus and, subsequently, tested on speech segments from another corpus.

Statistical analysis

Statistical analyses were performed using the statistical program R, version 2.12.2 (R Foundation for Statistical Computing). For each strategy, we computed the accuracy scores as well positive and negative values (PPV and NPV), specificity and sensitivity. To assess the statistical significance of the recognition scores from both the one-to-one and the cross-over conditions, we used binomial tests comparing our experimental values and chance (=0.5 in the different data set). We also calculated the 95% confidence intervals (95%CI).

Results

One-to-one condition

Table 3 summarized the best classifier metrics that were obtained by machine learning (SVM) with training and testing in the same language. As evidenced by the accuracy scores, we obtained moderate-to-excellent classification results. All the results were significantly different from chance ($p<1x10^{-10}$), showing that classifying with SVM the emotional vs. the non-emotional components of IDS was feasible for all the languages (English, Hebrew, Italian, Portuguese, French) based on the acoustic characteristics of the speech. In Hebrew, it was also possible to do so with the two different speakers available (fathers and mothers). Also despite the fact that in Italian the database included both mothers and fathers, the results remained above chance levels. Figure 2 graphically summarizes the accuracy performance according to the number of features that were automatically extracted by the languages in the one-to-one condition (here, SVM tested the same language as was used during training).

Cross-over condition

Table 4 shows the matrix of IDS emotional component accuracy scores that was obtained by machine learning for different learning and testing strategies. The diagonal corresponds to the accuracy metrics that were obtained by machine learning (SVM) in the one-to-one condition (training and testing in the same

language). As evidenced by the accuracy scores, we obtained mild-to-good classification results when the testing was performed in a different language than training. The best scores were obtained when testing Hebrew-speaking fathers and training in English-speaking mothers (acc=0.84; 95%CI: 0.82-1) or Italian-speaking caregivers (acc=0.81; 95%CI: 0.80-1). The lowest scores were obtained when testing Italian-speaking caregivers after training in Hebrew-speaking mothers (acc=0.53; 95%CI: 0.52-1) or after training in French-speaking mothers (acc=0.54; 95%CI:0.54-1). However, all the results were significantly different from chance ($p<1x10^{-10}$), showing that although cross-over conditions alter the level of accuracy scores that are achieved, they still classified emotional vs. non-emotional components of IDS above chance levels.

Discussion

Given that all the testing results were significant and above the chance level, the study shows that some common acoustic features are sufficient to determine what emotional IDS is. We summarized in table 5 the 3 main hypotheses of the study according to the main results. As evidenced by the one-to-one condition results that are summarized through the accuracy scores in the diagonal of table 2, there was no exception to this hypothesis. We did find the best accuracy scores in the one-to-one condition, showing that SVM machine learning was adapted for our classification needs. In addition, as shown in figure 2, best classification results are obtained when many acoustic descriptors are used, confirming what we have already shown in Italian caregivers [8]. In other words, acoustic descriptors that convey emotion in IDS are numerous and they cannot be summarized by the fundamental frequency F0 [6,37].

Regarding the second hypothesis, which is the core of cross-cultural comparison, we found that accuracy scores of testing another language than the one that was used during training were always significantly different form chance. Also in 90% of the cases (27/30), they were lower than those in the one-to-one condition. This is greater than the chance level. The three exceptions occurred when testing Hebrew-speaking fathers after training in English, Italian and Portuguese (green squares in table 4). The fact that testing Hebrew-speaking fathers yielded such high performance warrants several interpretations. First, the recordings in this database were excellent, as a specific microphone was used for each partner, including the father [32]. Second, the quality of fatherese was particularly high, as evidenced by the one-to-one condition accuracy scores for Hebrew-speaking fathers that

represented the best results (see Table 1). Third, although this might have been obtained by chance, we cannot exclude the possibility that the context of the recruitment through advertising for the purpose of research increased the quality and motivation of the fathers in this sample.

The third hypothesis was challenged by Hebrew speakers, as we had two independent databases for fathers and mothers. Upon testing the Hebrew-speaking fathers after training the Hebrew-speaking mothers, we found poorer results than in the Hebrew-speaking fathers in the one-to-one condition. When testing Hebrew-speaking mothers after training Hebrew-speaking fathers, we found poorer results than in the Hebrew mothers in the one-to-one condition (yellow squares in table 4). However, classification results remained above chance level. We conclude that common acoustical characteristics exist to classify the emotional component of IDS whoever the speaker. The results found in the Italian database in which mothers and fathers are mixed also support this view. In all cross-over conditions involving the Italian database, we found significant classification results above chance levels.

We are aware that the current results have numerous limitations. First, the databases were obtained from diverse sources and recording qualities that might have affected the IDS classifier metrics. Second, we cannot exclude that the different contexts of early interaction to record IDS (mainly experimental vs. natural contexts) affected the way parents produced IDS. For instance, the fact that, in the Hebrew database, parents know they will have to be unresponsive to the child at some point (Still face experiment) may have affected the way to use or not the emotional component of IDS. Third, we only tested hypothesis 3, which compared mothers vs. fathers in Hebrew. In Italian, in a previous study investigating parents' adaptation to infant psychopathology through home movies, we had the opportunity to develop another IDS classifier based on a Gaussian mixture model that was optimized for Italian mothers and Italian fathers separately and that could achieve accuracy scores above 0.8 [5,8]. Fourth, the age of the infants varied across datasets (4 and 18 months), which might have influenced the quality of the emotional IDS [6,7]. Finally, contrasts across corpora are not only contrasts in languages, but also contrasts between different situations. Therefore, we cannot exclude the possibility that the results might not be generalizable. Research is ongoing to collect natural interactive recording in different languages to control for these possible biases.

We conclude that the automated classification of emotional and non-emotional components of IDS is possible based on the acoustic characteristics regardless

of the language. The results found in the cross-over condition support the hypothesis that the emotional component shares similar acoustic characteristics across languages. The results also support the importance of acoustic features in conveying social signals from caregivers to infants during early interactions.

Author contribution

Conceived and designed the experiments: CSG, MC, DC. Performed the experiments: JMC, CSG, EP, JX. Analyzed the data: CSG, JMC, ZG, DC. Contributed reagents/materials/analysis tools: EP, RF, LO, FM, SV, JMC. Wrote the paper: EP, CSG, DC, MC. Revised the final version: All authors

Funding

The study was supported by the Agence Nationale de la Recherche (ANR-12-SAMA-006) and the Groupement de Recherche en Psychiatrie (GDR-3557). It was partially performed in the Labex SMART (ANR- 11-LABX-65), which is supported by French state funds and managed by the ANR in the Investissements d'Avenir program under reference ANR-11-IDEX-0004-02. The sponsors had no involvement in the study design, data analysis, or interpretation of the results.

Competing interests: The authors declare that they have no conflict of interest.

References

1. Falk D. Findings our tongues: mothers, infants and the origin of language. Basic Books, New York,2009.
2. Saxton, M. What's in a name? Coming to terms with the child's linguistic environment. J Child Lang, 2008; 35(3), 677-686.
3. Ferguson, C. A. Baby Talk in Six Languages. American Anthropologist, 1964; 66(6_PART2), 103-114.
4. Soderstrom, M. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. Developmental Review 2007; 27: 501-532.
5. Cohen D, Cassel RS, Saint-Georges C, Mahdhaoui A, Laznik Mc, Apicella F, Muratori P, Maestro S, Muratori F, Chetouani M. Do motherese prosody and fathers' commitment facilitate social interaction in infants who will later develop autism? PlosONE 2013; 8(5): e61402
6. Saint-Georges C, Chetouani M, Cassel Rs, Apicella F, Mahdhaoui A, Muratori P, Laznik Mc, Cohen D. Motherese, an emotion- and interaction-based process, affects infants' cognitive development. PlosONE 2013; 8: e78103
7. Spinelli M, Fasolo M, Mes man J . Does prosody make the difference? A meta-

analysis on relations between prosodic aspects of infant-directed speech and infant outcomes. Dev Rev, 2017; 44: 1-18.
8. Mahdhaoui A, Chetouani M, Cassel RS, Saint-Georges C, Parlato E, Laznik MC, et al. Computerized home video detection for motherese may help to study impaired interaction between infants who become autistic and their parents. Inter J Methods Psychiatr Res 2011; 20(1):e6-e18.
9. Papoušek M, Papoušek H, Symmes D. The meanings of melodies in motherese in tone and stress languages. Infant Behav Dev 1991; 14: 415-440.
10. Biben, M., Symmes, D., & Bernhards, D. Contour variables in vocal communication between squirrel monkey mothers and infants. Dev Psychobiol, 1989; 22(6): 617-631.
11. Snowdon, C. T., & Teie, D. Affective responses in tamarins elicited by species-specific music. Biol Lett 2010; 6(1), 30-32.
12. Falk, D. Prelinguistic evolution in early hominins: whence motherese? Behav Brain Sci 2004; 27(4), 491-503; discussion 503-483.
13. Fernald, A. Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In Barkow/Cosmides/Tooby, The Adapted Mind, 1992; pages 391-428.
14. Snow CE, Ferguson CA Talking to children. Cambridge, UK: Cambridge University Press, 1977.
15. Pye C. Quiché Mayan speech to children. J Child Lang. 1986; 13(1): 85-100.
16. Matsuda YT, Ueno K, Cheng K, Konishi Y, Mazuka R, Okanoya K. Auditory observation of infant-directed speech by mothers: Experience dependent interaction between language and emotion in the basal ganglia. Frontiers in Human Neuroscience 2014; 8: e907.
17. Vincarelli A, Pantic M, Bourlard H. Social signal processing: Survey of an emerging domain.Image and Vision Computing 2009; 27: 1743–1759.
18. Ringeval F, Demouy J, Szaszák G, Chetouani M, Robel L, Xavier J, Cohen D, Plaza M. Automatic Intonation Recognition for the Prosodic Assessment of Language Impaired Children. IEEE Transactions on Audio, Speech and Language Processing 2011; doi:10.1109/TASL.2010.2090147
19. Demouy J, Plaza M, Xavier J, Ringeval F, Chetouani M, Périsse D, Chauvin D, Viaux S, Golse B, Cohen D, Robel L. Differential language markers of pathology in Autism, Pervasive Developmental Disorder Not Otherwise Specified and Specific Language Impairment Research in Autism Spectrum Disorders 2011; 5: 1402-1412.
20. Girard JM, Cohn JF. Automated audiovisual depression analysis. Current Opinion in Psychology, 2015; 4: 75-79.
21. Avril M, Leclère C, Viaux S, Michelet S, Achard C, Missonnier S, Keren M, Cohen D, Chetouani M. Social signal processing for studying parent-infant interaction. Frontiers in Psychology. 2014; 5: 1-14.
22. Leclere C, Avril M, Viaux-Salevon S, Bodeau N, Achard C, Missonnier S, Keren M, Feldman R, Chetouani M, Cohen D. Interaction and behaviour imaging: a novel method to measure mother-infant interaction using video 3D reconstruction. Translational Psychiatry. 2016; 6: e816.
23. Hammal Z, Cohn JF, Messinger DS. Head movement dynamics during play and

perturbed mother-infant interaction. IEEE Transactions on Affective Computing, 2015; 6(4): 361-370.
24. Liu HM, Tsao FM, Kuhl PK. Acoustic analysis of lexical tone in Mandarin infant-directed speech. Dev Psychol 2007; 43: 912-917.
25. Fais L, Kajikawa S, Amano S, Werker JF. Now you hear it, now you don't: vowel devoicing in Japanese infant-directed speech. J Child Lang England 2010; 37: 319-340.
26. Reilly JS, Bellugi U. Competition on the face: affect and language in ASL motherese. J Child Lang 1996; 23: 219-239
27. Hoff E, Tian C. Socio-economic status and cultural influences on language. J Commun Disord 2015; 38: 271-278.
28. Hoff-Ginsberg E. Mother-child conversation in different social classes and communicative settings. Child Dev 1991; 62: 782-796.
29. Rowe ML. Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. J Child Lang 2008; 35: 185-205.
30. Feldman R, Singer M, Zagoory O. Touch attenuates infants' physiological reactivity to stress. Developmental Science 2010; 13 271–278.
31. Weisman O, Delaherche E, Rondeau M, Chetouani M, Cohen D, Feldman R. Oxytocin shapes parental motion during father-infant interaction. Biol. Lett. 2013; 9: 20130828
32. Weisman O, Zagoory-Sharon O, Feldman R. Oxytocin administration to parent enhances infant physiological and behavioral readiness for social engagement. Biol Psychiatry 2012; 72: 982–989.
33. Slaney M, McRoberts G. Baby Ears: a recognition system for affective vocalizations. Proceeding of the 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle, WA, May 12-15, 1998. Copyright 1998, IEEE.
34. Desjardins V, Foki J, Chauveau D, Delmas JF. Analyse statistique de la communication par le système perceptif d'un bébé (de 3 à 9 mois) avec sa mère. 2008. <hal-00324170>
35. Schuller B, Steidl E, Batliner A, Vinciarelli A, Scherer K, Ringeval F, Chetouani M, Weninger F, Eyben F, Marchi E, Mortillaro M, Salamin H, Polychroniou A, Valente F, Kim S. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. Proc. Interspeech 2013; 5 pages.
36. Eyben F, Wollmer M, Schuller B. Opensmile: the Munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM international conference on Multimedia. 2010; Pages 1459-1562.
37. Chetouani M, Mahdhaoui A, Ringeval F. Time-Scale Feature Extractions for Emotional Speech Characterization. Cog Comp 2009; 2:194-201.

Table 1: Overview of the infant-directed speech database					
Source	Language	Speaker	Infant age (months)	Setting	Utterance (N)
Slaney and McRoberts, 1998	English	Mother (N=6)	10 to 18	Audio-recorded free play interaction	2322
Feldman et al., 2010	Hebrew	Mother (N=53)	4 to 7	Video-recorded free play interaction during a still face experiment	3490
Weisman et al., 2012	Hebrew	Father (N=32)	3 and 6	Video-recorded free play interaction during a still face experiment	1994
Cohen et al., 2013	Italian	Mothers and fathers (N=30)	3 to 18	Home movies of natural social interaction from infants with typical development and infants who will later develop autism	35504
Unpublished	Portuguese	Mother (N=13)	3 to 12	Video-recorded free play interaction	5734
Desjardins et al., 2008	French	Mother (N=5)	6 to 10	Video-recorded free play interaction	3622

Table 2. OpenSMILE Low-Level Descriptors (LLD) and functionals (statistical, polynomial regression coefficients, and transformations) that can be applied to LDD

Feature Group	Description	Functionals to be applied
Waveform	Zero-Crossings, Extremes, DC	Extremes (Extreme values, positions, and ranges) Means (Arithmetic, quadratic, geometric) Moments (Std. dev., variance, kurtosis, skewness) Percentiles (Percentiles and percentile ranges) Regression (Linear and quad. approximation coefficients, regression err., and centroid) Peaks (Number of peaks, mean peak distance, mean peak amplitude) Segments (Number of segments based on delta thresholding, mean segment length) Sample values (Values of the contour at configurable relative positions) Times/durations (Up- and down-level times, rise/fall times, duration) Onsets (Number of onsets, relative position of first/last on-/offset) DCT (Coefficients of the Discrete Cosine Transformation (DCT) Zero-Crossings (Zero-crossing rate, Mean-crossing rate)
Signal energy	Root Mean-Square & logarithmic	
Loudness	Intensity & approx. Loudness	
FFT spectrum	Phase, magnitude (lin, dB, dBA)	
ACF, Cepstrum	Autocorrelation and Cepstrum	
Mel/Bark spectr.	Bands 0-*Nmel*	
Semitone spectr.	FFT based and filter based	
Cepstral	Cepstral features, e.g. MFCC, PLPCC	
Pitch	*F*0 via ACF and SHS methods Probability of Voicing	
Voice Quality	HNR, Jitter, Shimmer	
LPC	LPC coeff., reflect. coeff., residual Line spectral pairs (LSP)	
Auditory	Auditory spectra and PLP coeff.	
Formants	Centre frequencies and bandwidths	
Spectral	Energy in *N* user-defined bands, multiple roll-off points, centroid, entropy, flux, and rel. pos. of max./min.	
Tonal	CHROMA, CENS, CHROMA based features	

Adapted from Eyben et al. [36]

Table 3: Best classifier metrics obtained by machine learning (SVM) with training and testing in the same language to classify emotional vs. non-emotional components of IDS

	Accuracy	PPV	NPV	Specificity	Sensitivity	P value*	95%CI
English	0.81	0.69	0.84	0.61	0.88	$<10^{-10}$	0.79-1
Hebrew M	0.69	0.74	0.71	0.76	0.69	$<10^{-10}$	0.68-1
Hebrew F	0.91	0.92	0.91	0.92	0.91	$<10^{-10}$	0.9-1
Italian	0.63	0.59	0.59	0.58	0.6	$<10^{-10}$	0.62-1
Portuguese	0.64	0.64	0.67	0.61	0.7	$<10^{-10}$	0.63-1
French	0.70	0.73	0.62	0.5	0.82	$<10^{-10}$	0.68-1

M=mother; F=father; PPV=positive predictive value; NPV=negative predictive value. Accuracy chance level equal to 0.5. 95%CI=95% Confidence Interval

*The binomial test compared classification results with a random guess [here, equal to 0.5].

Table 4: Emotional component of IDS accuracy scores that were obtained by machine learning with training in one language and testing in another language

		TRAINING					
		English (N=2322)	Hebrew M (N=3490)	Hebrew F (N=1994)	Italian (N=35504)	Portugese (N=5734)	French (N=3622)
		Open source data base[1]	Experimental database[2]	Experimental database[3]	Home movie data base[4]	Clinical database	Clinical database[5]
TESTING	English	0.81	0.58	0.67	0.66	0.62	0.62
	Hebrew M	0.64	0.69	0.62	0.62	0.63	0.64
	Hebrew F	0.84	0.67	0.91	0.81	0.74	0.70
	Italian	0.58	0.53	0.57	0.63	0.57	0.54
	Portugese	0.58	0.57	0.56	0.57	0.64	0.58
	French	0.58	0.57	0.57	0.58	0.57	0.70

M=mother; F=father; [1]Baby-Ear [34]; [2]Details available in [30]; [3]Details available in [32]; [4]Details available in [5,24]; [5]Details available in [35]

Table 5. Summary of the main findings according to the study hypotheses		
	Hypothesis	Main findings
H1	Testing the same language as was used during training allowed for the best emotional vs. non-emotional IDS classifier metrics to be obtained.	For all the languages, the best accuracy scores during testing were found in the one-to-one condition (orange diagonal in table 2).
H2	Testing another language than was used during training should yield lower metrics in general that are still above the chance level, which allowed for the generalization capabilities of the best emotional vs. non-emotional IDS classifier to be obtained.	For all the possible cross-over combinations (N=30), the accuracy scores of testing another language than was used during training were lower than in the one-to-one condition in 27/30 (90%) combinations.
H3	Changing the gender of the speakers during training and testing in a database should yield lower metrics in general that are still above the chance level.	When testing the Hebrew-speaking fathers after training the Hebrew mothers, we found poorer results than in the Hebrew fathers in the one-to-one condition. The opposite was also true.

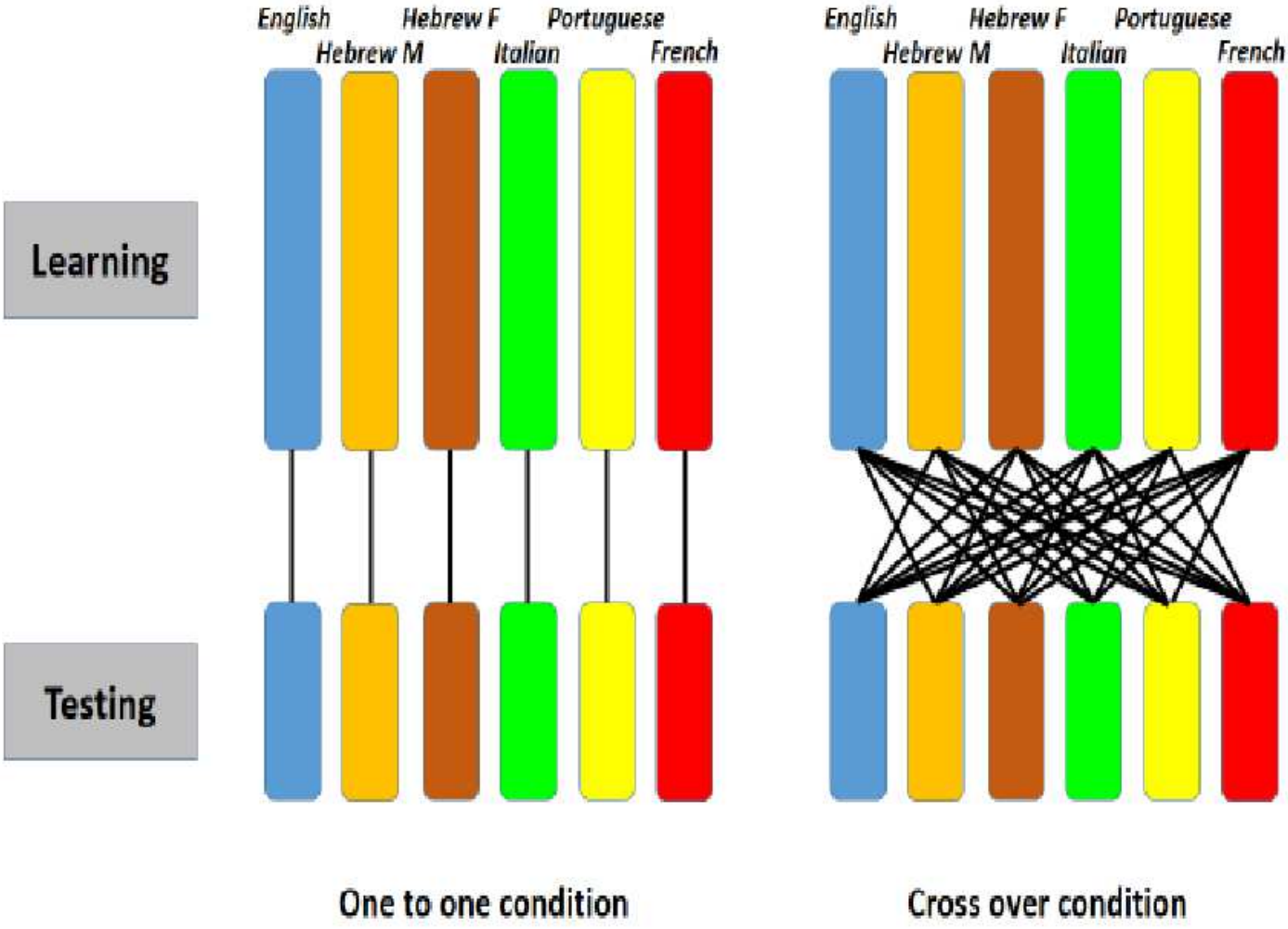

Figure 1. Conditions of SVM learning and testing according to the language corpus

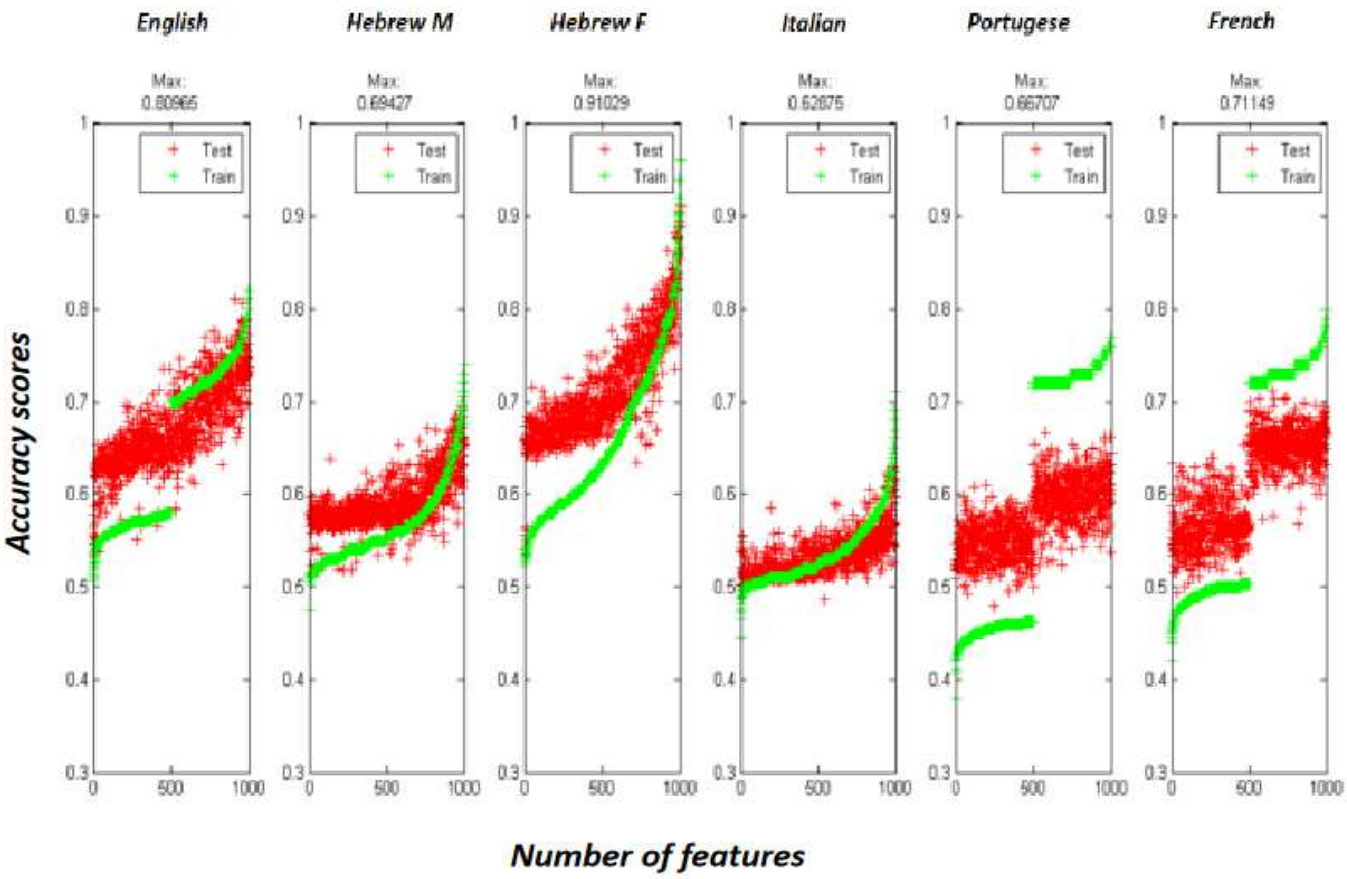

Figure 2. Accuracy performances according to the number of features automatically extracted by the languages (here, SVM tested the same language as was used during training).

For the ease of presentation, the figures are based on the first 1000 features contributing for each classification.