

Running head: BABY CRY

Automatic identification of expiratory and inspiratory sounds in baby cry audio recordings

Jean-Julien Aucouturier

Temple University, Japan Campus &
ERATO, Japan Science and Technology Agency &
Biolinguistics Laboratory, Brain Science Institute,
RIKEN, Wako, Japan

Yulri Nonaka

Biolinguistics Laboratory, Brain Science Institute,
RIKEN, Wako, Japan

Kentaro Katahira

ERATO, Japan Science and Technology Agency

Kazuo Okanoya

ERATO, Japan Science and Technology Agency &
Biolinguistics Laboratory, Brain Science Institute,
RIKEN, Wako, Japan

Abstract

We describe an automatic technique to identify expiration and inspiration phases from the audio recording of human baby cries. The algorithm uses a two-stage recognition architecture: first, spectral characteristics of both types of sounds are learned, then recognized using a support vector machine classifier. Second, the classifier decisions are converted into output probabilities and fed to a Viterbi decoder similarly to a hidden Markov model. The algorithm yields up to 95% classification precision (86% average), and we demonstrate its generalization ability over different babies, different ages and vocalization contexts. The technique can be used to quantify expiration duration, count the crying rate and other time-related characteristics of baby crying for diagnosis and research purposes.

Automatic identification of expiratory and inspiratory sounds in baby cry audio recordings

During the first year of life, infants gradually acquire language by transitioning from gestures and pre-linguistic vocalizations to referential speech. Baby cries even in the neonatal period are signals of intriguing acoustical complexity. Their communication bases are still poorly understood. On the one hand, the cry appear to be a graded signal that reflects certain states of the infant such as need or pain (Soltis, 2004). On the other hand, increasing evidence shows that the cry is a categorical signal that conveys particular type of needs (Okanoya, 2007). Besides its use for language research, acoustic analysis of baby cry is also an important tool to diagnose a variety of medical conditions, such as asphyxia (Pearce & Taylor, 1993) or sudden infant death syndrome (Corwin & Lester, 1995).

A frequently used measure to evaluate infant crying is its fundamental frequency (F0), which was shown to depend on age and body weight (Baeck & Souza, 2006). Another typical diagnosis and analysis methodology is to compute the energy distribution in the sound's spectrogram (Pearce & Taylor, 1993). Both are average measurements which can be utilized on the whole audio signal, without prior segmentation into expiratory and inspiratory stages. However, recent research has started to recognize the importance of the temporal structure of the cry. As for adults, expiratory and inspiratory infant vocalizations have well differentiated characteristics (Orlikoff, Baken, & Kraus, 1997). The expiratory rate may correlate with certain physiological states such as pain (Soltis, 2004). Within expiratory phases, several types of pitch contours can be observed, which has been hypothesized to be protolinguistic context markers (Nonaka, Katahira, Shiba, & Okanoya, 2008).

All such analyses require prior segmentation and indexation of the recorded audio material into phases of expiration and inspiration. Manually annotating large amounts of

data is time-consuming and error-prone: Orlikoff et al. (Orlikoff et al., 1997) uses simultaneous measurements of airflow data to assist the process. Even when large amounts of data are available, analysis is typically restrained to a limited range, e.g. recordings in the sole 2nd month of age in (Nonaka et al., 2008). Such small sample sizes, coupled with a typically large inter-subject variability, severely limit claims for statistical significance of any observed results (Nonaka et al., 2008), generating frequent contradictory findings (Baeck & Souza, 2006). For diagnosis purposes too, manual annotation is impractical and impairs the immediacy, safety and cost of any subsequent automatic analysis.

The following article presents a fully automatic technique to segment and identify expiration and inspiration phases from the audio recording of human baby cries. The algorithm uses a two-stage pattern recognition architecture: first, spectral characteristics of both types of sounds are learned from a small corpus of annotated data, then generalized to unseen data using support vector machines (SVM) and gaussian mixture model (GMM) classifiers. Second, the classifier decisions are converted into output probabilities and fed to a Viterbi decoder, similarly to a hidden Markov model. The technique is found to generalize well over a variety of subjects and crying contexts, thus to be applicable for a wide range of subsequent acoustic analysis. Fully automatic, it can process large amounts of data and enables longitudinal studies over development time, individual and contextual differences that were impractical so far. As an example, we describe a straightforward application of the technique to measure and compare the length of expiratory vocalizations for 14 babies, over 12-months, in 4 different crying contexts.

Algorithm

Audio Processing

Figure 1 shows a spectrogram of an audio recording of four successive expirations and inspirations by a 1-month-old baby, interspaced with short silence phases. It appears

that inspired phases are characterized by breathier, less voiced sounds than expirations. This is consistent with inspiratory airflow being higher than in expiration for both infants and adults (Orlikoff et al., 1997), as well as the anatomic configuration of the vocal folds favoring pitched phonation in an aggressive direction (Grau & Robb, 1995); hence, noise-sensitive audio characteristics such as the rate of zero-crossings of the waveform (L. R. Rabiner & Schafer, 1978) or the flatness of its power spectrum (Peeters, 2003) are natural candidates to detect inspiratory sounds. However, careful inspection of the data shows that expiration phases also often contain noisy episodes, such as growls, which are easily confused with inspiration with such features, yielding an insatisfactorily high amount of false positives (e.g. time 640 in Figure 1 has a higher zero-crossing rate than neighboring inspiratory phases).

Therefore, we opt for a generic feature description of the audio spectrum: audio signals are reduced to mono, cut into series of overlapping frames using a 20ms Hanning window with 10ms hop size, and each frame is converted to a set of 13 Mel-Frequency Cepstrum Coefficients (L. Rabiner & Juang, 1993). The cepstrum is the inverse Fourier transform of the log-spectrum $\log \mathcal{S}$.

$$c_n = \frac{1}{2\pi} \int_{\omega=-\pi}^{\omega=\pi} \log \mathcal{S}(\omega) \exp j\omega n d\omega \quad (1)$$

We call mel-cepstrum the cepstrum computed after a non-linear frequency warping onto a perceptual frequency scale, the Mel-frequency scale. The c_n in Eq. 1 are called Mel frequency cepstrum coefficients (MFCCs). Cepstrum coefficients provide a low-dimensional, smoothed version of the log spectrum, and thus are a good and compact representation of the spectral shape. They are widely used as feature for speech recognition, and have also proved useful in e.g. musical instrument recognition (Essid, Richard, & David, 2006) and environmental sound monitoring (Aucouturier, Defreville, &

Pachet, 2007). In this work, we use the first $n = 13$ MFCCs, including 0-th order (which is a correlate of the frame’s RMS energy).

First stage: Classifier

We use a two-stage classification algorithm. In the first stage, a decision is made for each successive frame of audio data (20ms) individually. Each frame is classified as either “expiratory” (EX), “inspiratory” (IN) sound or “silence” (SI), the latter being a waste-basket category including glimpses of background sound perceived in between vocalisations. To this aim, we investigate here two types of pattern recognition algorithms: Gaussian mixture models and Support-vector machines.

A Gaussian mixture model (GMM) is a technique to estimate a probability density, which it models as the weighted sum of \mathcal{M} simpler Gaussian densities, called components or states of the mixture. ((Bishop, 1995)):

$$p(x_i) = \sum_{m=1}^{m=\mathcal{M}} \pi_m \mathcal{N}(x_i, \mu_m, \Sigma_m) \quad (2)$$

where x_i is the feature vector observed at time i , \mathcal{N} is a Gaussian pdf with mean μ_m , covariance matrix Σ_m , and π_m is a mixture coefficient (also called state prior probability). Given a set of observations (e.g. many example frames of “expiratory” sounds, encoded as MFCCs), the parameters of the GMM are learned so as to maximize the probability of the model given the observations, using e.g. the classic E-M algorithm ((Bishop, 1995)). We train a different GMM for each of the 3 classes {EX,IN,SI}. Given an unknown frame of audio, each class GMM produces an estimate of the probability of the data $p(x_i)$, and the GMM with highest probability is chosen as the most likely class (a strategy known as “maximum likelihood” decision). The whole process is implemented in Matlab using the Netlab toolbox¹.

A support-vector machine (SVM) does not estimate class probabilities, but rather

class boundaries. It does so by mapping its training vectors are mapped into a higher dimensional space using a function θ . It then finds a linear separating hyperplane with the maximal margin between classes (in order to guarantee it is generalizable to unseen data) in this higher dimensional space. A computational trick makes the training of a SVM possible without specifying the mapping function θ , but only the vector product $\mathcal{K}(x_i; x_j) = \theta(x_i)^T \theta(x_j)$, called the kernel function. In this work, we conduct all experiments with the so-called radial basis function (RBF), defined as

$$\mathcal{K}(x_i; x_j) = \exp -\gamma \|x_i - x_j\|^2, \gamma > 0 \quad (3)$$

where x_i, x_j are two training vectors expressed in the space of MFCCs. The classification accuracy of a SVM largely depends on the choice of the kernel parameter γ and a penalty parameter C which controls the amount to which deviations are tolerated. In standard practice (Hsu, Chang, & Lin, 2003), we estimated these two parameters by grid search, using a 10-fold cross-validation scheme on our training data (see below): the best values were $\gamma=C=1$. Although this is not required for classification, we additionally convert the decision function value for each frame into posterior probabilities, using Platt's expression:

$$p(class/x_i) = 1 / \exp A * f(x) + B \quad (4)$$

where x_i is a training vector and A, B are two scaling parameters estimated by Maximum Likelihood over the training set (Platt, 1999). The whole process is implemented using a Matlab wrapper for the LibSVM library ².

Second stage: Viterbi decoding

Both techniques above (GMM and SVM) make a class decision between {EX,IN,SI} for each audio frame, independently. While the 3 classes are sufficiently clustered and

distinct in the MFCC feature space to warrant decent classification at this stage, additional information can be exploited in the temporal context of the frames. Figure 2-top shows a manual annotation into {EX,IN,SI} categories of a typical 3-mn bout of crying of a 2-month old baby. We observe that, accordingly to intuition, expiratory vocalizations are nearly always followed by an inspiratory phase; both types of phases tends to last for several seconds (i.e. several hundred frames), and expiration tends to be longer; silence phases typically only occur between EX and IN, or between IN and EX (and not, say, in between two EX phases); silence is more frequent after an inspiratory phase than after an expiratory phase. Figure 2-middle shows a typical output of a SVM classifier trained to do frame-per-frame decisions on the same data. We observe that, while the decisions are globally in accordance with manual classifications, their arrangement in time often disagree with the above patterns. Nothing prevents the algorithm to label the audio as a series of short expiratory bursts without intermediate inspirations, or rapid infra-second alternations between EX and IN, both or which are at best unlikely if not physiologically impossible.

To avoid such mistakes, we add a second stage of processing to the frame-based classification, using the Viterbi algorithm. The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence (or path) of hidden states that results in a sequence of observed events (L. Rabiner & Juang, 1993). It does so by combining likelihood information on each frame with transition probabilities between states: a frame classified with high probability as e.g. IN may be assigned a lower probability if the transition to IN is unlikely at this position in the sequence (e.g. because we have just started a phase of EX frames, and the transition to IN would be unrealistically fast). At each new frame, the algorithm examines all possible paths leading to this frame and only keep the one most likely. This is implemented by recursion, as a classical dynamic programming problem. Figure 2-bottom shows the best sequence of

categories found by Viterbi on the same data as above. We observe that many of the unrealistic mistakes of the frame-per-frame classifier were corrected.

Viterbi algorithms is typically used in hidden Markov models (HMM) to decode the most probable sequence of states once individual state likelihood is given for each frame (L. Rabiner & Juang, 1993), e.g. in automatic speech recognition. Therefore, the two-stage architecture used here can be understood as a hidden Markov model, where state output distributions {EX,IN,SI} are modelled using a GMM or SVM-based probabilistic model. Typical HMM training (using e.g. the Baum-Welsh algorithm) infers both state output distributions and state-wise transition probabilities in a concurrent manner. However in this work, we have training data on the sequence of hidden states, hence each state’s probability distribution can be trained independently and simply integrated in a “manual” HMM.

Results

Data

We collected audio recordings of baby cries, for 14 distinct babies. Recordings were taken in each baby’s home environment in the Tokyo area, voluntarily by the baby caregiver. Recordings were regularly spaced throughout the babies’ first year of life, starting at a few days’ old. For each recording, context (as understood by the caregiver) was annotated as one of a set of fixed labels including “hungry”, “lonely”, etc. All recordings are 44,1kHz mono audio, captured by consumer digital audio recorders. Recordings were then edited and segmented into 30-second files, each representing a bout of crying. Of all files, we annotated manually recordings of 3 babies (id:044,050 and 051) for 3 contexts (“hungry”, “sleepy”, “pee”). Annotations identify the beginning and end point of each audible expiratory and inspiratory sound, and were made using CCNY’s Sound Analysis Pro³. Details of the annotated data, including total number of expiratory

and inspiratory phases, are found in Table 1. On the whole, the annotated data used in this study amounts to 1h47' of audio and more than 8000 sound events.

Performance

We compare here the performance of 6 algorithmic variants for our 3-class classification problem {EX,IN,SI}:

- SVM classifier, with and without Viterbi correction
- GMM classifier, trained with 5 gaussian components, with and without Viterbi
- GMM classifier, 20 gaussian components, with and without Viterbi

We compared performances on 16 different training/testing datasets, combining data from one or several babies, and one or several contexts (see tabulation in Table2 for more details).

All algorithms are tested using the same evaluation scheme: 10-fold cross-validation on a dataset of 10,000 frames extracted randomly. More precisely, Viterbi correction requires that the training and testing data be organized in bouts of consecutive frames. Therefore, we generated the dataset by randomizing the files, but preserving time order within files, so that data frames within all testing folds were consecutive in time.

As seen earlier, SVM parameters are taken to be $\gamma=C=1$ for all folds and all datasets, but output probability parameters are estimated separately for each fold during training. GMMs are trained with diagonal covariance matrices, using 20 iterations of k-means initialization and 50 iterations of EM optimization. Transition matrices used for Viterbi correction are constructed anew from the training data at each fold.

Table 2 reports cross-validation performance for all algorithms and all datasets. The mean performance over all settings is $M=82.7\%$ ($SD=3.6\%$) without Viterbi correction and $M=86.1\%$ ($SD=3.1\%$) with Viterbi. Best accuracies for each baby frequently exceed 90% with a maximum of 93.5% (050-hungry). Variations of accuracy over different babies

are of the same amplitude as variations over different contexts: the mean accuracies for all viterbi-corrected tests, for all contexts are 85.5% (baby:44), 88.7% (baby:50) and 84.9% (baby:51); the mean accuracy for all viterbi-corrected tests, for all babies are 87.5% (context:hunger), 85,5% (context:pee), 85,5% (context:sleepy).

Performance of the frame-based SVM is c. 2% higher than frame-based GMMs for both 5 and 20 gaussian components (M=83.9% against 82.3% and 81.8% resp.). For all algorithms, Viterbi correction improves accuracy by a mean 3.5%. Interestingly, Viterbi contributes more accuracy to the frame-based GMM (M=+4.1% for 5-GMM, M=+4.6% for 20-GMM) than to the frame-based SVMs (M=+1.7%). This can be due to the way SVMs outputs are scaled to look like probabilities: SVM are not statistical estimators per-se, and their discriminative power results in very contrasted probabilities for each class (the likelihood of the winning class being orders of magnitudes bigger than the other classes'). While this is a good property for frame-based decisions, this reduces leeway for optimization at the Viterbi stage: even highly unlikely state transitions cannot counterbalance such large likelihood differences between classes. On the contrary, GMMs do not guarantee good discrimination between classes, but rather goodness of fit to each class distribution. The resulting class likelihoods are typically more balanced, with winning classes only marginally superior to their alternatives. While this allows for more noise on a frame-per-frame basis (hence reduced accuracy), this seems to give more flexibility for Viterbi correction to overweight maximum likelihood decisions and possibly explains its greater benefit than for SVMs.

On the whole, the best performance is reached for Viterbi-corrected GMMs, with no effect of the number of GMM components (86.3% for 5-GMM, 86.4% for 20-GMM), topping at 93.5%.

Generalizability over developmental time

Acoustic properties of baby vocalizations change with development time. For example, mean F0 largely oscillates in the first 6 months of life, with a significant decrease from 0 to 15 days of age (Baeck & Souza, 2006). Such variations are likely to influence the classification accuracy of algorithms such as we present here, if e.g. an acoustic model is trained using 1-month-old data and tested over 6-month-old data. To evaluate such developmental effects, we conduct a multi-slice cross validation where we train models at early development time, and test them on increasingly older datasets. More precisely, for a given baby and given context, we divide all recordings into 10 slices, spanning the development scale evenly (each slice corresponds roughly to one-month worth of audio). We train models on slice 1 (first month of age) and test them on each successive slice from 2 to 10. For each configuration (e.g. slice 1 vs slice 3), we conduct 10 evaluation trials, each with a random training set and testing set of identical size (5000 frames). Figure 3 shows the evolution of testing accuracy of a SVM model (without Viterbi correction) on the 044-hungry dataset. Testing accuracy steadily decreases as the time interval between training and testing data increases. top performance is reached for testing in the training slice (94.2% accuracy). Testing after 8 or 10 months results in degradation as big as -35% absolute. The profile can be roughly fitted with a linear curve, with slope $a=-3.8\%$ per month, $R=0.79$. Similar trends are observed for all babies, all contexts in our dataset. In other words, our models generalize poorly with time.

This has several implications. Practically, to ensure good generalization, training data needs to incorporate recordings at varied developmental ages. A possible heuristic is to sample the typical development range at regular time intervals. The performance reported in Table 2 took no account of development age and selected training data randomly over the whole time range; the resulting accuracies were in the high range of the developmental profile (85%-90%). Small variations across babies and contexts in Table 2

suggest that training over recordings at several developmental stages is more important than training over recordings of many babies and contexts. This does not mean that there is no acoustic correlate of e.g. cry context in baby vocalizations (see (Nonaka et al., 2008; Soltis, 2004) and Section below). However, this suggests that context may not affect the respective spectral properties of expiratory and inspiratory phonation, possibly rather their sequence in time or their pitch contour.

From a technical point of view, such developmental effects act as an additional hidden variable over our observation space, which can be subjected to statistical inference and thus be learned. While our present models do not incorporate development time, but only local time in the form of frame transitions, it would be feasible to learn the developmental changes of the class distributions as well as their local properties, thus adding a third layer of time processing to the models presented here. Whether such models would warrant better generalization is an open research question.

More generally, the analysis of the generalization ability of such pattern recognition algorithms over individuals, contexts and development time has potential to shed new lights on development studies, addressing such questions as “do individuals increasingly differentiate from one another with development time?” or “do fully developed vocalization repertoires subsume vocalizations at earlier stages?”. This will be the focus of our future work.

Applications

We present here one possible application of the technique to study the effect of context on the average duration of expiratory and inspiratory vocalizations of baby cries from 0 to 3 months of age. We analyze vocalizations from 12 babies, over 4 different contexts (hungry, pee, sleepy and lonely). We train one of the best performing classifiers identified in Table 2 (5-GMM with Viterbi correction) over a random selection of 10,000

annotated frames for baby id:044,050,051, and use it to segment all other recordings into 3 phases {EX,IN,SI}. Table 3 shows the statistics of mean duration, cumulated over all babies and broken down by context and age (0-3 MOA).

We observe that the mean duration of expiratory phases increases with development age, in all vocalization contexts. On average, +76ms from M0-M1 and +81ms from M1-M2. This increase may correspond to physiological factors such as increased lung capacity, as already noted by (Baeck & Souza, 2006). On the same period, no clear pattern can be observed for inspiratory duration.

More interestingly, the standard deviation of expiratory durations between contexts increases with development age, leading to significant differences between contexts at M2 and M3. Figure 4 illustrates this gradual divergence of mean duration according to context, with increasing development age. This divergence may correspond to a greater degree of respiratory control, and possibly some voluntary differentiation of vocal behaviour according to situation. The two series of mean expiratory durations, averaged over each of the 12 babies, are statistically significantly different between contexts PEE and SLEEPY at both M2 and M3, according to a non-parametric Kruskal-Wallis test of ANOVA (e.g. $F(1,17)=4.26$, $p=0.039$ for M2). In contrast, none of the context pairs show any significant difference at either M1 (at best, HUN vs SLE: $F(1,15)=0.15$, $p=0.696$) or M0 (at best, PEE vs SLE, $F(1,10)=1.04$, $p=0.308$). This is in accordance with previous studies using manually annotated data (Nonaka et al., 2008). These studies however lacked statistical validation over many individuals, something only made possible by the automated technique introduced in this work.

Conclusion

We described a fully automatic technique to segment and identify expiratory and inspiratory phases from the audio recording of human baby cries. One innovation of the

algorithm is to process the decisions of a typical frame-per-frame classifier with a subsequent Viterbi module, to ensure realistic temporal statistics of the output. In its best configuration, the technique yields 86.4% average precision, topping at 93.5%. One advantage of the technique presented here is that it is fully automatic: it can thus process large amounts of data and enables longitudinal studies over development time, individual and contextual differences that were impractical so far. As an example, we computed the mean duration of expiratory and inspiratory phases over a dataset of 12 babies, from 0 to 3 MOA, and 4 vocalization contexts. We found that the cry duration was significantly longer for the Pee context than for the Sleepy context at 2 and 3 months of age but not earlier. This demonstrates the potential of our method as a pre-processing for both language acquisition studies and diagnosis.

References

- Aucouturier, J.-J., Defreville, B., & Pachet, F. (2007). The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *Journal of the Acoustical Society of America*, *122*(2), 881-91.
- Baeck, H. E., & Souza, M. Nogueira de. (2006). Longitudinal study of the fundamental frequency of hunger cries along the first 6 months of healthy babies. *Journal of Voice*, *21*(5), 551-559.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford Press.
- Corwin, M., & Lester, B. (1995). Newborn acoustic cry characteristics of infants subsequently dying of sudden infant death syndrome. *Pediatrics*, *96*, 7377.
- Essid, S., Richard, G., & David, B. (2006). Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Transactions on Speech and Audio Processing*.
- Grau, S. M., & Robb, M. P. (1995). Acoustic correlates of inspiratory phonation during infant cry. *Journal of Speech and Hearing Research*, *38* (2), 373-382.
- Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). *A practical guide to support vector classification*. Technical report, Department of Computer Science, National Taiwan University.
- Nonaka, Y., Katahira, K., Shiba, R., & Okanoya, K. (2008). Development of infant cry acoustics: a basis of musical and linguistic skills. In *Proceedings of 10th international conference on music perception and cognition, sapporo, japan*.
- Okanoya, K. (2007). Language evolution and an emergent property. *Current Opinion in Neurobiology*, *17*, 271-276.
- Orlikoff, R. F., Baken, R. J., & Kraus, D. H. (1997). Acoustic and physiologic characteristics of inspiratory phonation. *J. Acoust. Soc. Am.*, *102* (3), 1838-1845.
- Pearce, S., & Taylor, B. (1993). Energy distribution in the spectrograms of the cries of

- normal and birth asphyxiated infants. *Physiol Meas.*, 14, 263268.
- Peeters, G. (2003). *A large set of audio features for sound description*. IRCAM Technical Report.
- Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*.
- Rabiner, L., & Juang, B. (1993). *Fundamentals of speech recognition*. Prentice-Hall.
- Rabiner, L. R., & Schafer, R. W. (Eds.). (1978). *Digital processing of speech signals*. Englewood Cliffs, New Jersey: Prentice Hall.
- Soltis, J. (2004). The signal functions of early infant crying. *Behavioral and Brain Sciences*, 27, 443-490.

Notes

¹<http://www.ncrg.aston.ac.uk/netlab/index.php>

²<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

³http://ofer.sci.ccny.cuny.edu:2001/html/sound_analysis.html

Table 1

Details of the annotated data used for training and testing in this work. We used recordings from 3 different babies, in 3 different contexts (as identified by the caregiver). On the whole, we collected about 1h50' of audio, amounting to c. 5800 expiratory phases, 2600 inpiratory phases and 8420 phases of non-vocalized, background noise and silence (each phase consisting of many frames)

id	Context	Data		Spanning		# Phases		
		# Files	Length	From	To	EX	IN	NO
44	Hungry	44	22'	0m12d	11m10d	1255	894	2149
	Pee	22	11'	0m8d	3m5d	776	697	1471
	Sleepy	18	9'	1m2d	13m28d	431	223	654
50	Hungry	47	23'30	0m13d	6m7d	1117	116	1233
	Pee	14	7'	0m17d	3m23d	303	33	336
	Sleepy	19	9'30	1m26d	6m23d	445	45	490
51	Hungry	38	19'	0m4d	6m3d	1125	439	1565
	Pee	1	0'30	2m28d	-	26	26	52
	Sleepy	11	5'30	0m4d	3m9d	337	134	470
Total		214	1h47'	0m4d	13m28d	5815	2607	8420

Table 2

Cross-validation accuracy of 3-class classification over 6 algorithmic variants and 16 datasets, each tested with a 10-fold scheme over 10,000 random frames. Best results are achieved for Viterbi-corrected GMMs, with mean accuracy 86.4% and maximum: 93.5%. Viterbi correction contributes c. 3.5% extra accuracy, and helps GMMs more than SVMs. Dataset 51-pee () was excluded because of too few files.*

id	context	SVM		GMM5		GMM20	
		frm	vtb	frm	vtb	frm	vtb
44	HUN	84.6%	88.0%	83.4%	87.0%	83.4%	85.9%
	PEE	83.0%	85.8%	84.6%	87.1%	78.0%	84.4%
	SLE	86.5%	87.8%	79.0%	84.8%	80.2%	85.8%
	ALL	79.5%	81.9%	78.7%	83.2%	79.2%	84.4%
50	HUN	87.3%	86.5%	83.0%	88.1%	85.6%	93.5%
	PEE	88.8%	91.0%	85.0%	88.7%	87.2%	91.7%
	SLE	78.6%	80.1%	82.0%	85.5%	85.7%	89.5%
	ALL	87.0%	86.4%	88.5%	91.6%	89.4%	92.3%
51	HUN	85.1%	87.3%	81.4%	83.8%	83.0%	86.0%
	PEE(*)	-	-	-	-	-	-
	SLE	84.1%	85.4%	79.0%	83.7%	76.8%	83.3%
	ALL	85.8%	87.0%	82.6%	86.5%	76.9%	81.9%
ALL	HUN	85.1%	86.7%	79.4%	84.8%	78.8%	82.7
	PEE	83.0%	85.5%	86.9%	89.7%	85.2%	88.1%
	SLE	75.7%	78.8%	80.7%	84.6%	76.1%	81.5%
	ALL	83.9%	86.2%	80.4%	85.7%	81.4%	85.2%

Table 3

Statistics of the duration of expiratory and inspiratory phases, analysed automatically, cumulated over all 12 babies in our dataset. Statistically significant differences of expiratory durations between contexts are observed in older babies (M2 and M3) (marked with asterix)

M	Context	Duration (ms)	
		Expiration	Inspiration
0	Hungry	M=421,SD=429	M=303,SD=320
	Pee	M=462,SD=275	M=255,SD=220
	Sleepy	M=415,SD=427	M=192,SD=185
	LON	M=409,SD=327	M=261,SD=164
	ALL	M=435,SD=380	M=278,SD=278
1	Hungry	M=507,SD=475	M=265,SD=325
	Pee	M=510,SD=460	M=231,SD=195
	Sleepy	M=522,SD=509	M=245,SD=240
	LON	M=496,SD=386	M=204,SD=122
	ALL	M=511,SD=475	M=248,SD=265
2	Hungry	M=696,SD=675	M=284,SD=287
	Pee	* M=571,SD=496	M=271,SD=249
	Sleepy	* M=453,SD=436	M=380,SD=387
	LON	M=525,SD=681	M=172,SD=108
	ALL	M=592,SD=590	M=315,SD=323
3	Hungry	M=640, SD=590	M=261,SD=355
	Pee	M=598,SD=462	M=330,SD=279
	Sleepy	M=481,SD=374	M=493,SD=969
	LON	M=579,SD=549	M=558,SD=368
	ALL	M=608,SD=557	M=370,SD=628

Figure Captions

Figure 1. Top: Spectrogram of an audio recording of four successive expirations and inspirations by a 1-month-old baby. Inspiratory stages (here labelled manually, and marked by dotted borders) are characterized by breathier, less voiced sounds than expirations. Bottom: Zero-crossing rate (ZCR) of the same audio data, measured on successive 50-ms time windows. While inspiratory stages are marked by high ZCR values, these are easily confused with occasional noisy episodes in expiratory vocalizations, such as time 640.

Figure 2. Top: manual annotation (into expiratory, inspiratory and silence) of a 3-minute bout of crying by a 2-month old baby. Middle: Frame-per-frame classification of the same stimuli using a typical SVM classifier: precision is satisfying but the arrangement in time of the 3 possible phases is unrealistic, with many successive, short phases. Bottom: Filtering of the frame-per-frame decisions using the Viterbi algorithm.

Figure 3. Testing accuracy of a SVM non Viterbi-corrected classifier trained on the first month of age of baby:044, and tested on 10 successive age slices. Testing accuracy steadily decreases as the time interval between training and testing data increases, down to -35% absolute.

Figure 4. Mean duration of expiratory and inspiratory vocalizations, averaged over 12 babies, for different contexts (hungry, pee, sleepy and lonely) and developmental age. Statistically significant differences of expiratory durations between contexts are observed in older babies. Standard error bars omitted for clarity (see Table3)







