*INTERSPEECH 2016*
September 8–12, 2016, San Francisco, USA

# Manual Versus Automated: The Challenging Routine of Infant Vocalisation Segmentation in Home Videos to Study Neuro(mal)development

*Florian B. Pokorny*[1,2,3], *Robert Peharz*[1,3], *Wolfgang Roth*[4], *Matthias Zöhrer*[4], *Franz Pernkopf*[3,4], *Peter B. Marschik*[1,3,5], *Björn W. Schuller*[6,7]

[1]Research Unit iDN – interdisciplinary Developmental Neuroscience, Institute of Physiology, Center for Physiological Medicine, Medical University of Graz, Austria
[2]Machine Intelligence & Signal Processing group, Technische Universität München, Germany
[3]Brain, Ears & Eyes – Pattern Recognition Initiative (BEE-PRI), BioTechMed-Graz, Austria
[4]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
[5]Centre of Neurodevelopmental Disorders (KIND), Karolinska Institutet, Stockholm, Sweden
[6]Chair of Complex & Intelligent Systems, University of Passau, Germany
[7]Machine Learning Group, Department of Computing, Imperial College London, UK
florian.pokorny@medunigraz.at

## Abstract

In recent years, voice activity detection has been a highly researched field, due to its importance as input stage in many real-world applications. Automated detection of vocalisations in the very first year of life is still a stepchild of this field. On our quest defining acoustic parameters in pre-linguistic vocalisations as markers for neuro(mal)development, we are confronted with the challenge of manually segmenting and annotating hours of variable quality home video material for sequences of infant voice/vocalisations. While in total our corpus comprises video footage of typically developing infants and infants with various neurodevelopmental disorders of more than a year running time, only a small proportion has been processed so far. This calls for automated assistance tools for detecting and/or segmenting infant utterances from real-live video recordings. In this paper, we investigated several approaches of infant voice detection and segmentation, including a rule-based voice activity detector, hidden Markov models with Gaussian mixture observation models, support vector machines, and random forests. Results indicate that the applied methods could be well applied in a semi-automated retrieval of infant utterances from highly non-standardised footage. At the same time, our results show that, a fully automated approach for this problem is yet to come.
**Index Terms**: voice activity detection, infant vocalisation, home video database, retrospective audio-video analysis

## 1. Introduction

Voice activity detection, i. e., the attempt to automatically extract segments of speech from background noise, is an important requirement for the front-end of a number of real-world speech processing systems, such as automatic speech recognition systems [1, 2]. At the expense of traditional rule-based voice activity detection approaches (e. g., [3, 4, 5, 6, 7, 8]), machine-learning-based approaches have become increasingly popular in recent years (e. g., [9, 10, 11, 12, 13, 14, 15]). A main research focus has been put on the improvement of voice activity detection for speaker-independent applications under real-world settings [16, 17].

However, due to limited fields of application a limited number of studies dealt with the automatic detection/segmentation of infant voice (e. g., [18]).

For the last 20 years, our research has – inter alia – focussed on speech-language phenomena. We have been studying typically developing (TD) infants, infants with brain injury, and infants with neurodevelopmental disorders characterised by a mean age of diagnosis in or beyond toddlerhood (conditions of interest hereafter; COI: e. g., autism spectrum disorder, ASD; Rett syndrome, RTT; or fragile X syndrome, FXS). Our overall aim is to define behavioural biomarkers – especially in the motor and speech-language domain – to facilitate earlier identification. However, the diagnosis beyond toddlerhood in combination with a low prevalence of most of our COI (rare genetic disorders) hampers the implementation of comprehensive prospective studies. Therefore, we have been collecting audio-video data of the above mentioned conditions to build a decent corpus with home video material. The retrospective analyses of home video material of COI-infants in the prodromal period, i. e. in the first year of life, were based on vocalisation sequences manually segmented from the videos clips. The efforts of this approach call for a reliably performing infant voice activity detector that would (i) facilitate the time-consuming segmentation process and (ii) constitute the essential input stage for an automated vocalisation-based tool for the early detection of maldevelopment.

In the following, we introduce our research database and investigate different voice activity detection/vocalisation segmentation approaches.

## 2. Methods

### 2.1. Database

In this study, infant voice activity detection experiments were carried out far beyond studio conditions on a real-world database actively used in medical/neuro-physiological research – the Graz University Audiovisual Research Database for the Interdisciplinary Analysis of Neuro(mal)development (GUARDIAN).



http://dx.doi.org/10.21437/Interspeech.2016-1341

#### 2.1.1. Material

GUARDIAN has been built up over 15 years and currently comprises home video material with a total running time of more than a year. The corpus comprises videos of TD infants, infants with brain injury, and infants later diagnosed with COI, i. e., to a large extent rare genetic disorders and ASD. Videos were recorded by the infants' parents during typical family situations (playing, feeding, bathing, etc.) and during special family events (birthday parties, Christmas eve, etc). At the time of recording, the parents of infants later diagnosed with COI were not aware of their child's medical condition. Information on the infants' exact age in months was known or reconstructed for each scene. The videos were provided by the families after having received a COI-diagnosis for the purpose of retrospective audio-video analysis to define early markers of various COI.

GUARDIAN's audio-video material is characterised by inhomogeneity in terms of signal quality/original audio-video format/codec, used recording device, infant's nationality/family language, recording setting (camera angle, number of persons present in different scenes, etc.), recording location (indoor and outdoor), or date of recording (referring to recording year).

Due to the great effort needed for data pre-processing in context of video-based retrospective speech-language analysis including scene selection and behaviour coding/annotation, only a fraction of our whole database has been fully annotated so far, thus, prepared for scientific analysis. For this study, we used a representative subset of GUARDIAN comprising more than 20 hours of manually annotated audio sequences that were extracted from home videos diverging in a number of meta-parameters: (a) the earliest videos were shot in 1988, the latest ones in 2009; (b) the original record carriers varied from analogue (e. g., Super 8, Hi8, VHS, SVHS, Video 8) to digital (e. g., DV tape, DVD) formats; (c) the material contains TD infants, infants later diagnosed with ASD or RTT, and one infant later diagnosed with FXS; and (d) the videos were provided by families from four different nationalities (Austria, Germany, Italy, and United Kingdom) with three different mother tongues/family languages (German, Italian, English). All recordings in the dataset originate from the infants' respective second half year of life, i. e., from months 7 to 12.

#### 2.1.2. Annotation

Audio-video data were manually annotated for infant vocalisations in terms of setting start and stop markers using the video coding system Noldus Observer XT. Vocalisations were defined as utterances correlating to vocal breathing groups [19]. Vegetative sounds (e. g., breathing sounds, sneezes, hiccups, smacking sounds) were not annotated. Furthermore, we did not annotate vocalisations not produced by the participating infant with absolute certainty (e. g., in situations with the participating infant and other infants present in the scene) and vocalisations incomplete due to a jump cut in the video. The searching for relevant vocalisations as well as raw segmentation was done by three female and four male research assistants following detailed instructions and three training sessions by the first author. Prior to inclusion in this study, the first author verified each pre-selected vocalisation and performed the fine segmentation. The dataset used for this study comprised a total of 4 903 annotated prelinguistic vocalisations with a mean length of 1.72s ($\pm$ 1.41s standard deviation). Each included video contained at least one annotated vocalisation. Detailed information on the dataset such as meta-parameter-specific numbers of annotated vocalisations is given in Table 1. For matters of speech-language analysis vocalisations were further annotated for vocalisation types not considered in this study.

Table 1: *Number of infants (inf), length (l) of available audio-video material in format hours(h):minutes(m):seconds(s), and number of manually segmented vocalisations (voc) with respect to COI, gender (GEN), and nationality/mother tongue (NAT/L1). ASD = autism spectrum disorder; AT = Austria; eng = English; f = female; FXS = fragile X syndrome; ger = German; IT = Italy; ita = Italian; m = male; RTT = Rett syndrome; TD = typical development; UK = United Kingdom.*

| COI | GEN | NAT/L1 | $\#_{\text{inf}}$ | l [h:m:s] | $\#_{\text{voc}}$ |
|---|---|---|---|---|---|
| ASD | m | IT/ita | 13 | 02:59:51 | 696 |
| FXS | m | AT/ger | 1 | 00:17:05 | 87 |
| RTT | f | AT/ger | 2 | 02:18:02 | 454 |
| RTT | f | DE/ger | 2 | 05:54:12 | 1745 |
| RTT | f | UK/eng | 5 | 02:08:30 | 386 |
| TD | f | AT/ger | 4 | 04:55:27 | 1044 |
| TD | m | AT/ger | 5 | 02:18:39 | 491 |
| | | $\Sigma$ | 32 | 20:51:50 | 4903 |

### 2.2. Voice activity detection

In this study, we investigated different voice activity detection/vocalisation segmentation approaches (2.2.1–2.2.4) in order to identify and discuss their strengths and weaknesses with respect to our specific data and application area. Voice activity detection/vocalisation segmentation was carried out on the basis of the audio tracks extracted from the home videos in format 16 kHz/16 bit/1 channel/PCM. The manually annotated infant vocalisations represented the reference segments. All detection/segmentation approaches were implemented to operate on frames of size 25 ms and a step size of 10 ms.

#### 2.2.1. cVadV1

As an example for a standard rule-based voice activity detector we selected the cVadV1 component of the open source feature extraction toolkit openSMILE [20] in its current release [21]. In this detector the decision for voice versus non-voice is simply made on the basis of fuzzy scores related to deviations from the mean long-term trajectories of energy, line spectral frequencies, and Mel spectra [22].

#### 2.2.2. HMMs

In our hidden Markov model (HMM) approach we used two states indicating presence/absence of infant speech in the corresponding frame. We used a uniform prior and a transition probability of $0.005$ for both states, i. e., a-priori the model stays with probability $0.995$ in each of the two states, encouraging a strong state blocking which we also observed in the training data (see Section 3). We trained Gaussian mixture models (GMMs) as observation models, using the expectation maximisation (EM) algorithm and cross-validating the number of components. State prediction was performed using the Viterbi algorithm. We first used 100 features, comprising pitch [23], 13 Mel-frequency cepstral coefficients (MFCCs), 13 perceptual linear predictive coefficients (PLPs), 13 Rasta-PLPs and the 60 features used in the CLEAR 2006/2007 challenges [24]. We normalised the data to zero-mean and unit variance. For train-

ing, we used a stratified sub-set of the frames, using all positive frames and the same number of negative frames. Since using all features lead to poor results, we first applied a greedy feature forward selection, i. e., we trained a classifier on each individual feature and selected the one leading to the best acoustic event detection accuracy (AED-ACC; see Section 3.1) on the validation set. We iterated this process, holding the selected features fixed; in this way we selected eight features, since the validation performance degraded thereafter. Furthermore, using these eight selected features, we applied discriminative GMMs as observation models by means of large-margin training [25]. To the HMM using GMMs trained with EM we refer as $\text{HMM}_{\text{gen}}$; to the HMM with discriminative GMMs we refer as $\text{HMM}_{\text{disc}}$.

#### *2.2.3. SVMs*

For training support vector machines (SVMs) we used all 100 features described in Section 2.2.2 and reduced the training set to 125 000 samples per class, as the training time of SVMs increased drastically with the amount of data. We used a Gaussian kernel and cross-validated the kernel width $\gamma \in \{2^{-10}, \ldots, 2^{10}\}$ and the trade-off factor $C \in \{2^{-10}, \ldots, 2^{10}\}$, where AED-ACC was optimised on the validation set. Prediction with SVM was done frame-wise, subsequently using a median filter of length 15 to smooth the system's output.

#### *2.2.4. Random forests*

We trained Random Forests (RFs) [26] using super-vectors [27] of the 100 features described in Section 2.2.2. We cross-validated the number of trees $T \in \{50, 100, 200, 300\}$, the maximal depth $D \in \{5, 10, 15, 20\}$, and the minimum number of samples per leaf $M \in \{1, 10, 100\}$, where AED-ACC was optimised on the validation set. As for SVMs, the prediction was done frame-wise, subsequently using a median filter of length 15.

## 3. Experiments

Performance evaluation of all investigated voice activity detection/vocalisation segmentation approaches was carried out on the basis of a subset (test set) of our dataset. The remaining subset (training set) was used for training (first half of audio frames) and validating (second half of audio frames) the presented machine-learning-based approaches, i. e., HMMs, SVMs, and RFs. Partitioning was done speaker/infant-independently in a way that two third of infants per diagnosis, gender, and family language were part of the training set and one third part of the test set. For example, vocalisations of two third of girls later diagnosed with RTT from English speaking families were part of the training set, etc. For the final configuration we further considered the number of vocalisations per infant to obtain a roughly two third to one third distribution in absolute vocalisation number between training and test sets. Vocalisations of the single individual later diagnosed with FXS were assigned to the training set to avoid testing on an unknown COI. The detailed partitioning into training and test set is given in Table 2.

The bases for performance evaluation of each investigated approach were the raw binary output vectors with '0' indicating 'no infant voice present' and '1' indicating 'infant voice present' in a frame (of 25 ms). Before calculating any evaluation measures, a moving median filter with a filter length of 15 frames was applied to the raw output vectors (for SVMs and RFs already done as part of the classification model). The filter length was chosen according to the shortest reference vocalisation occurring in the training set (150 ms). By filtering, (i) detected voice periods shorter than 8 frames were eliminated, and (ii) two detected voice segments interrupted by a non-voice period shorter than 8 frames were concatenated to one single segment.

Table 2: *Dataset partitioning into training and test set with specification of respective number of infants (inf) and number of vocalisations (voc) in dependence of the infants' COI, gender (GEN), nationality (NAT), and mother tongue/family language (L1). ASD = autism spectrum disorder; AT = Austria; DE = Germany; eng = English; f = female; FXS = fragile X syndrome; ger = German; IT = Italy; ita = Italian; m = male; RTT = Rett syndrome; TD = typical development; UK = United Kingdom.*

| | | | Training | | Test | |
|---|---|---|---|---|---|---|
| COI | GEN | NAT/L1 | $\#_{\text{inf}}$ | $\#_{\text{voc}}$ | $\#_{\text{inf}}$ | $\#_{\text{voc}}$ |
| ASD | m | IT/ita | 9 | 424 | 5 | 272 |
| FXS | m | AT/ger | 1 | 87 | - | - |
| RTT | f | AT∪DE/ger | 3 | 1836 | 1 | 363 |
| RTT | f | UK/eng | 3 | 262 | 2 | 124 |
| TD | f | AT/ger | 3 | 702 | 1 | 342 |
| TD | m | AT/ger | 3 | 339 | 2 | 152 |
| | | $\Sigma$ | 22 | 3650 | 11 | 1253 |

### 3.1. Measures

We evaluated the selected approaches for two different scopes of application. On the one hand, we were interested in a frame-based voice detection/segmentation accuracy as the basis for a fully automated vocalisation analysis system. In this case, we compared the filtered binary output vector of the voice activity detector with the binary reference vector (ground truth annotation) frame-by-frame and assigned and counted true positives (TPs), false positives (FPs), and false negatives (FNs). On the other hand, we evaluated the selected approaches for vocalisation-based accuracy in sense of detecting the (rough) location of an acoustic event/a vocalisation. According to [28], a reference vocalisation was considered to be correctly detected (TP), if the centre of a segment proposed by a detection approach was situated within the boundaries of the reference segment, or vice versa (i. e., the center of the reference segment was situated within the boundaries of a segment proposed by the detector). Figure 1 exemplifies the procedure of assigning TPs, FPs, and FNs.

For both frame-based and vocalisation-based performance evaluation we calculated precision and recall, as well as the acoustic event detection accuracy (AED-ACC $\equiv$ F-measure) and the acoustic event error rate (AEER) according to [28] given in Equations 1–4.

$$precision = \frac{\#\,TPs}{\#\,TPs + \#\,FPs} \tag{1}$$

$$recall = \frac{\#\,TPs}{\#\,TPs + \#\,FNs} \tag{2}$$

$$AED-ACC = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

$$AEER = \frac{\#\,FNs + \#\,FPs}{\#\,TPs + \#\,FNs} \tag{4}$$

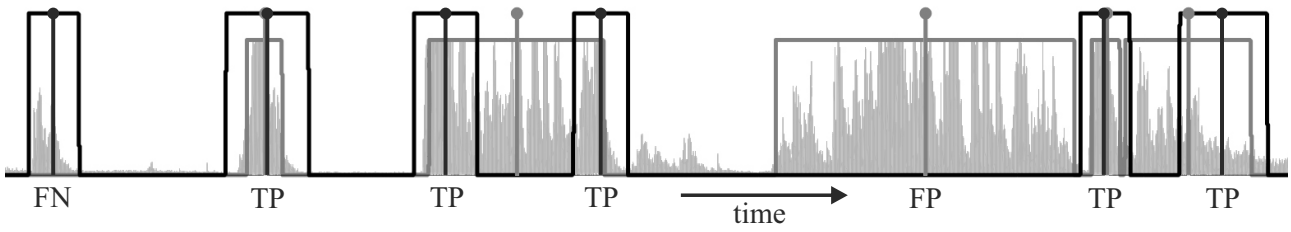


Figure 1: *Absolute value of waveform of a sample audio sequence (grey) with manually segmented reference vocalisations (black frames) with segment centres (black vertical lines) and hypothetically detected/segmented vocalisations by a voice activity detection/segmentation approach (grey frames) with segment centres (grey vertical lines) and assignment to either true positive (TP), false positive (FP), and false negative (FN) as the basis for vocalisation-based performance evaluation.*

Table 3: *Frame-based and vocalisation-based voice activity detection/vocalisation segmentation performance of the investigated approaches. (Precision, recall, and AED-ACC are rounded to three decimal points. AEER is rounded to integers.)*

| Measure | cVadV1 | $HMM_{gen}$ | $HMM_{disc}$ | SVM | RF |
|---|---|---|---|---|---|
| Frame-based evaluation | | | | | |
| Precision | 0.148 | 0.181 | 0.204 | 0.194 | 0.109 |
| Recall | 0.587 | 0.661 | 0.599 | 0.557 | 0.789 |
| AED-ACC | 0.236 | **0.284** | 0.305 | 0.288 | 0.192 |
| AEER | 90295 | **74172** | 87597 | 96846 | 45994 |
| Vocalisation-based evaluation | | | | | |
| Precision | 0.105 | 0.172 | 0.255 | 0.111 | 0.070 |
| Recall | 0.886 | 0.886 | 0.740 | 0.876 | 0.956 |
| AED-ACC | 0.188 | **0.288** | **0.380** | 0.196 | 0.131 |
| AEER | 151 | **143** | **319** | 159 | 67 |

### 3.2. Results

Results for both the frame-based and the vocalisation-based evaluation of our investigated approaches are given in Table 3. For frame-based evaluation a good trade-off between AED-ACC and AEER was achieved when using the $HMM_{gen}$ approach. Vocalisation-based evaluated, our $HMM_{disc}$ approach reached the highest AED-ACC, but also the highest AEER. A good trade-off between AED-ACC and AEER was again achieved using $HMM_{gen}$.

## 4. Discussion

Our results varied in dependence of the used approach, but the maximum AED-ACCs achieved for frame-based and vocalisation-based evaluation were only 30.5% and 38.0%.

A factor dramatically degrading the performance of any voice activity detection/vocalisation segmentation approach in our study is the presence of numerous voice segments in the videos not produced by the respective participating infants (e. g., parental voice, voice from television or radio). Furthermore, in some scenes there are vocalisations produced by toddlers or infants other than the respective participating infant (e. g., vocalisations by elder brothers or sisters) or vocalisations produced by the participating infant but incomplete due to a jump cut causing FPs when being detected. For these and other reasons, the method of retrospective audio-video analysis based on home video recordings generally involves a number of limitations and risks, but by providing a unique window to 'look' or 'listen' to the past, at the moment, it is one of the best available means to study early phenomena in (rare) neurodevelopmental disorders with a late mean age of diagnosis, such as RTT [29] but also still ASD [30, 31, 32, 33].

Another issue concerning retrospective audio-video analysis is the absence of particular behaviours such as the occurrence of specific vocalisation types produced by an infant in an available dataset. Consequently, frequencies of behaviour occurrences should not be analysed on the basis of home video material. This brings a benefit for our endeavour of automatically detecting/segmenting infant vocalisations, because a missed vocalisation (FN) does not cause major drawbacks for further analysis, apart from a smaller dataset. Whereas, an incorrectly detected/segmented vocalisation (FP) would increase the expenditure of time for postprocessing or in case of fully automated vocalisation analysis/classification systems a FP would cause a bias.

## 5. Conclusions and Outlook

The manual segmentation of pre-linguistic vocalisations in the first year of life in variable quality home video material represents a challenging routine in our endeavour to retrospectively analyse speech-language development in individuals with neurodevelopmental disorders. Therefore, we introduced our non-standardised dataset of more than 20 hours of home video recordings including 4 903 annotated pre-linguistic infant vocalisations and investigated a number of voice activity detection/vocalisation segmentation approaches on the basis of our data. Results give reason to focus on the implementation of a semi-automated retrieval of vocalisations in near future. However, a fully automated approach is not feasible, yet.

Based on (semi-)automatically segmented infant vocalisations, in context of the analysis of speech-language development the automatic classification of pre-linguistic vocalisation types (e. g., [34, 35, 36]) should also be focused on in future.

## 6. Acknowledgements

The authors acknowledge funding from the Austrian Science Fund (FWF; P25241), the National Bank of Austria (OeNB; P16430), BioTechMed-Graz, the General Movements Trust, and the EU's H2020 Programme via RIA #688835 (DE-ENIGMA). Special thanks go to Andreas Kimmerle, Jorge Luis Moye, Sergio Roccabado, Iris Tomantschger, Adriana Villarroel, Diego Villarroel, and Claudia Zitta for their assistance in the vocalisation segmentation process. Moreover, thanks to Gunter Vogrinec for contributing to the dataset description. The authors express their gratitude to all parents who provided us with home video material for scientific analysis.

# 7. References


[1] J. Ramírez, J. C. Segura, J. M. Górriz, and L. García, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2177–2189, 2007.

[2] J. Ramírez, J. M. Górriz, and J. C. Segura, *Voice activity detection. Fundamentals and speech recognition system robustness*. INTECH Open Access Publisher, 2007.

[3] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[4] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 498–505, 2003.

[5] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.

[6] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *Signal Processing Letters, IEEE*, vol. 12, no. 10, pp. 689–692, 2005.

[7] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *Signal Processing, IEEE Transactions on*, vol. 54, no. 6, pp. 1965–1976, 2006.

[8] R. Tahmasbi and S. Rezaei, "A soft voice activity detection using GARCH filter and variance gamma distribution," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1129–1134, 2007.

[9] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *Signal Processing, 2002 6th International Conference on*, vol. 2. Beijing, China: IEEE, 2002, pp. 1124–1127.

[10] Q.-H. Jo, J.-H. Chang, J. Shin, and N. Kim, "Statistical model-based voice activity detection using support vector machine," *Signal Processing, IET*, vol. 3, no. 3, pp. 205–210, 2009.

[11] D. Cournapeau, S. Watanabe, A. Nakamura, and T. Kawahara, "Online unsupervised classification with model comparison in the variational Bayes framework for voice activity detection," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 6, pp. 1071–1083, 2010.

[12] J. W. Shin, J.-H. Chang, and N. S. Kim, "Voice activity detection based on statistical models and machine learning approaches," *Computer Speech & Language*, vol. 24, no. 3, pp. 515–530, 2010.

[13] J. Wu and X.-L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *Signal Processing Letters, IEEE*, vol. 18, no. 8, pp. 466–469, 2011.

[14] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2624–2633, 2011.

[15] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 697–710, 2013.

[16] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. Vancouver, Canada: IEEE, 2013, pp. 483–487.

[17] F. Germain, D. L. Sun, and G. J. Mysore, "Speaker and noise independent voice activity detection," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, ISCA. Lyon, France: ISCA, 2013, pp. 732–736.

[18] S. Yamamoto, Y. Yoshitomi, M. Tabuse, K. Kushida, and T. Asada, "Detection of baby voice and its application using speech recognition system and fundamental frequency analysis," in *Proc. of 10th WSEAS Int. Conf. on Applied Computer Science*, Iwate, Japan, 2010, pp. 341–345.

[19] M. P. Lynch, D. K. Oller, M. L. Steffens, and E. H. Buder, "Phrasing in prelinguistic vocalizations," *Developmental Psychobiology*, vol. 28, no. 1, pp. 3–25, 1995.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia, MM 2010*. Florence, Italy: ACM, October 2010, pp. 1459–1462.

[21] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia, MM 2013*. Barcelona, Spain: ACM, October 2013, pp. 835–838.

[22] F. Eyben, F. Weninger, M. Wöllmer, and B. Schuller, "openSource Media Interpretation by Large feature-space Extraction. Documentation Version 2.1," 2015.

[23] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis, Elsevier Science*, pp. 495–518, 1995.

[24] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," *Lecture Notes in Computer Science*, vol. 4122, pp. 311–322, 2007.

[25] F. Pernkopf and M. Wohlmayr, "Large margin learning of bayesian classifiers based on gaussian mixture models," in *ECML PKDD*, 2010, pp. 50–66.

[26] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[27] H. Phan, M. Maass, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 20–31, 2015.

[28] H. Phan, M. Maas, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 1, pp. 20–31, 2015.

[29] P. B. Marschik and C. Einspieler, "Methodological note: Video analysis of the early development of Rett syndrome – one method for many disciplines," *Developmental Neurorehabilitation*, vol. 14, no. 6, pp. 355–357, 2011.

[30] J. L. Adrien, P. Lenoir, J. Martineau, A. Perrot, L. Hameury, C. Larmande, and D. Sauvage, "Blind ratings of early symptoms of autism based upon family home movies," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 32, no. 3, pp. 617–626, 1993.

[31] E. R. Crais, L. R. Watson, G. T. Baranek, and J. S. Reznick, "Early identification of autism: How early can we go?" *Seminars in Speech and Language*, vol. 27, no. 3, pp. 143–160, 2006.

[32] R. Palomo, M. Belinchón, and S. Ozonoff, "Autism and family home movies: A comprehensive review," *Journal of Developmental & Behavioral Pediatrics*, vol. 27, no. 2, pp. 59–68, 2006.

[33] C. Saint-Georges, R. S. Cassel, D. Cohen, M. Chetouani, M.-C. Laznik, S. Maestro, and F. Muratori, "What studies of family home movies can teach us about autistic infants: A literature review," *Research in Autism Spectrum Disorders*, vol. 4, no. 3, pp. 355–366, 2010.

[34] D. K. Oller, "The emergence of the sounds of speech in infancy," *Child Phonology*, vol. 1, pp. 93–112, 1980.

[35] R. E. Stark, "Stages of speech development in the first year of life," *Child Phonology*, vol. 1, pp. 73–92, 1980.

[36] S. Nathani, D. J. Ertmer, and R. E. Stark, "Assessing vocal development in infants and toddlers," *Clinical Linguistics & Phonetics*, vol. 20, no. 5, pp. 351–369, 2006.