Automated Classification of Children's Linguistic versus Non-Linguistic Vocalisations

Zixing Zhang[1], *Alejandrina Cristia*[2], *Anne S. Warlaumont*[3], *Björn Schuller*[1,4]

[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2] LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, France
[3] Department of Communication, University of California, Los Angeles, USA
[4]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

zixing.zhang@imperial.ac.uk

Abstract

A key outstanding task for speech technology involves dealing with non-standard speakers, notably young children. Distinguishing children's linguistic from non-linguistic vocalisations is crucial for a number of applied and fundamental research goals, and yet there are few systems available for such a classification. This paper investigates two large-scale frame-level acoustic feature sets (eGeMAPS and ComParE16) followed by a dynamic model (GRU-RNN), and two kinds of derived static feature sets on the segment level (functional-based and Bag of Audio Words) combined with a static model (SVM), and automatically learnt representations directly from original raw voice signals by using an end-to-end system. These are applied to a large database of children's vocalisations (total N = 6,298) drawn from daylong recordings gathered in Namibia, Bolivia, and Vanuatu. Among these systems, the one implemented with GRU-RNN using ComParE16 features empirically performs best. We further identify promising paths of further research, including the application of a finer-grained classification of children's vocalisations onto these data, and the exploration of other feature systems.

Index Terms: infancy, linguistic vocalisations, babbling, crying, language acquisition, end-to-end, large-scale feature set, bag of audio words

1. Introduction

To be able to distinguish linguistic (e. g., speech and babbling [1, 2, 3, 4, 5, 6]) from non-linguistic (e. g., crying) sounds of children is important for at least three reasons. First, researchers and clinicians often need to study individual and group development of these different vocalisation types. For instance, a prior study [7] found that, from 10 to 48 months of age, American children's use of linguistic vocalisations increased steadily relative to their use of non-linguistic vocalisations. The study [7] also found that the rate of this growth was related to socioeconomic status and differed for children with and without autism. This type of research would be enhanced by being able to perform similar automated analyses using linguistically diverse datasets, and by testing whether the effects are robust to the particular classification algorithm used. Better linguistic versus non-linguistic child vocalisation classifiers could also eventually lead to improved segmentation procedures, which are particularly useful given the rise in use of day-long audio recordings (as those used in [7]). Finally, development of an open automated tool for classifying early human vocalisations as linguistic versus non-linguistic may also help scientists to gain a better understanding of the acoustic features that distinguish the two categories of sounds. Such information could potentially inform our understanding of the similarities and differences in the ways these sounds are produced [8]. Thus, it could further shed light on possible similarities and differences in the neural mechanisms of vocal tract control for linguistic vs reflexive sounds. This has been a topic of considerable interest to researchers interested in the evolution of human vocal communication (e. g., [9, 10, 11]).

To our knowledge, the only currently available system for automatically classifying human infant vocalisations as either linguistic/pre-linguistic or non-linguistic is the LENA™ system [12].[1] The LENA system uses a GMM-HMM approach [12] to identify child vocalisation periods within day-long home audio recordings, and to identify speech-related (i. e., linguistic) utterances and instances of cry, laugh, or vegetative (i. e., non-linguistic) sounds within those child vocalisation periods. The software is only available for use with audio recordings collected using the LENA audio recording hardware and is not open source.

Here we aim to contribute to the emerging literature that aims to classify children's vocalisations as a function of their linguistic properties, using a rather comprehensive approach, as follows. Two large-scale frame-level acoustic feature sets that are usually employed for other computational paralinguistics are extensively evaluated by either static models or dynamic models. For the static model, we selected the widely used Support Vector Machines (SVMs). To capture the segment-spanned information for static models, we conducted a functional-based approach and a bag-of-audio-words approach to extract the sequential low-level descriptors into a segment-level feature vector. For the dynamic model, we selected Recurrent Neural Networks (RNNs) equipped with Gated Recurrent Units (GRUs) due to their capability to capture long-term dependences. Finally, we investigated an end-to-end system to automatically learn representations from scratch instead of using the hand-engineered acoustic features.

2. Database Description

We built this database from three different datasets, gathered and annotated in the same way: In all cases, a small child wore a small portable recording device (USB, Olympus, or LENA recorders). One minute per hour was extracted and coded in Praat using a custom-written script (skipping the first 30 min-

[1]A number of studies have worked on automatic detection of emotion in infant vocalisation (e. g., [13, 14, 15, 16, 17]) or on automatic detection of different types of pre-linguistic infant vocalizations (e. g., [18, 19, 20]).

Table 1: *Data distribution over different partitions and categories of the selected database.*

partitions	ling.	non-ling.	$\sum$
training	2 267	280	2 547
development	1 604	212	1 816
test	1 605	330	1 935
$\sum$	5 476	822	6 298

utes of the recording to allow children to acclimate). One highly-proficient coder segmented and diarised the speech signal using both the spectrogram and auditory impression. Each child vocalisation was furthermore classified as linguistic (having at least a vowel) versus non-linguistic (typically, crying or laughing). Whining was considered linguistic as long as there was a recognisable vowel or some other such structure. For the present experiments, we further partitioned the samples via the subject-independent strategy into training, development, and test set, as shown in Table 1. Particularly, the development set is employed to optimise hyper-parameters of models.

The three datasets differed in terms of the population they involved. One portion was collected in a largely monolingual Yu'hoan-speaking village in Namibia [21]. Each of the 12 children (half girls; between 3 months and 3 years at initial intake) was recorded for between 2 and 5 consecutive days in a given visit, and the village was visited three times with 4 months intervals between the visits. The second portion was collected in a largely monolingual Tsimane-speaking village in Bolivia [22]. About half of the 27 children (9 girls; between 6 months and 5 years) were recorded twice. The third portion was gathered in several villages located in one island in Vanuatu [23]. None of the households was monolingual, with between 2 and 8 languages reportedly spoken (including Bislama and Venen Taut). The 13 children (7 girls between 5 and 37 months), were recorded once.

3. Automated Classification Systems

Traditional automatic speech recognition systems employ only (log)Mel filterbank energies, Mel Frequency Cepstral Coefficients (MFCCs), or others such features [24]. Following [25, 26], we preferred to consider much higher dimensional acoustic features for a more general intelligent speech analysis, namely the extended Geneva Minimalistic Acoustic Parameter Set (*eGeMAPS*) [27] and the one used for a series of INTERSPEECH Computational Paralinguistics Challenges (*ComParE16*) [26]. Table 2 displays the detailed information of the 130 frame-wise Low-Level Descriptors (LLDs) in the ComParE16 feature set, which contains not only commonly used features (MFCCs, f_0), but also other features of the same and other types. Note that the eGeMAPS, including 23 LLDs, can largely be considered as a core subset of ComParE16, selected according to brute-force empirical evaluations for computational paralinguistic tasks. More detailed information about eGeMAPS can be found in [27].

The LLDs represent the transitory acoustic patterns, but the identification of (non-)linguistic vocalisations benefits from longer-term temporal cues. For example, crying or laughing have alternations of ingressive and egressive air flow, and linguistic vocalisations include long non-canonical and muffled syllables as well as vocalisations with clear syllabic structure. To exploit this information, we can: i) transform the sequential LLD contours into the statistical features on the supra-segmental level due to the long-term essence of acoustic cues.

Table 2: *The* COMPARE *acoustic feature set includes 65 low-level descriptors (LLDs) of different types, as well as their first derivation (delta), resulting in 130 LLDs.*

4 energy related LLDs	**Group**
RMS energy, zero-crossing rate	Prosodic
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic
55 spectral LLDs	**Group**
MFCC 1–14	Cepstral
Psychoacoustic sharpness, harmonicity	Spectral
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	Spectral
Spectral energy 250–650 Hz, 1 k–4 kHz	Spectral
Spectral flux, centroid, entropy, slope	Spectral
Spectral roll-off point 0.25, 0.5, 0.75, 0.9	Spectral
Spectral variance, skewness, kurtosis	Spectral
6 frequency related LLDs	**Group**
f_0 (SHS and Viterbi smoothing)	Prosodic
Prob. of voicing	Voice quality
log. HNR, jitter (local and δ), shimmer (local)	Voice quality

By doing this, we can freely select the *static* models, such as SVMs, to analyse speech patterns, which are widely used for computational paralinguistics [26, 27]. We considered two approaches to derive the sequential frame-level LLDs into the segment-level vector, followed by SVM classification; ii) implement a *dynamic* model to learn the patterns directly from sequential LLDs. To this end, we used the GRU-RNNs; iii) exploit an *end-to-end* system that is able to automatically learn representations from the original raw voice signals, rather than using the hand-engineered LLDs. In the following, we introduce the four main automated classification systems, respectively.

3.1. Static Modelling with Functional-based Features

The functional-based approach has been frequently and successfully applied to these pattern recognition tasks, such as emotion recognition, which highly rely on long-term acoustic characteristics [28, 25]. Intuitively, this approach projects the temporal LLD contours onto a set of feature vectors with descriptive statistic functionals (see [26] and [27] for more detail). Mathematically, this can be written as follows:

$$\mathbf{z} = f([\mathbf{x}_i], i = 1, \ldots, T), \tag{1}$$

where $\mathbf{z}$ denotes the segment-level feature vector; $[\mathbf{x}_i]$ indicates the sequential frame-wise LLDs; T is the total frames of a given vocalisation; and f denotes the *functionals* (i. e., statistic information) that are applied per each LLD contour. Specifically, the functionals can include: extremes (minimum, maximum, ranges, etc.), mean (arithmetic, quadratic, geometric), moments (variance, skewness, kurtosis, etc.), percentiles (quantiles, ranges, etc.), peaks (number, distances, etc.), temporal variables (durations, positions, etc.), and regression (coefficients, error).

For our experiments, we applied several different functionals (see [27] for more details) to the eGeMAPS LLDs and the ComParE16 LLDs, which result in 88 and 6 373 dimensional feature vectors for the functional-based eGeMAPS and ComParE16 feature sets, respectively. Next, we submitted the functional-based feature vectors to SVMs for vocalisation classification.

3.2. Static Modelling with Bag-of-Audio-Words

Bag-of-Audio-Words (BoAWs) have been demonstrated to be alternative efficient representations for patterns found in rela-

tively long-term acoustic traits [29, 30].The extraction process involves three steps: i) *codebook generation* (which is done on the basis of extracted features); ii) *vector quantisation* (which takes into account both the codebook and the raw extracted features); and iii) *histogram construction* (which takes into account only the output of the vector quantisation step).

For linguistic analyses, the total number of context-words (codewords) is normally limited for a given language. In contrast, for acoustic analyses, the total number of audio-words (frame-wise LLDs) is numerous with an equal occurrence frequency of one, due to the fact that each frame-wise LLD normally has different values. To reduce the codebook size (S), a k-means clustering or a random sampling is conducted on the training set to generate the codewords for the codebook (C) [29, 30]. After that, a multi-assignment quantisation technique is executed to map each audio-word to the first N closest codewords (W) in the codebook, measured by Euclidean distance. Finally, a histogram is constructed by calculating the counts of occurrence of each codeword in all acoustic frames over one vocalisation segment. To minimise the effects relating to the length disparities of different vocalisations, a normalisation process is further undertaken over the histogram, to sum up all elements of $\mathbf{g}$ to one. Finally, we fed the BoAW representations to SVMs.

3.3. Dynamic GRU-RNNs Modelling

Different from the above two systems, this system directly models the variations of sequential acoustic features in a continuum. To do this, we selected the RNNs equipped with GRUs, which was initially proposed in [31]. Compared with other feedforward neural networks or classic RNN, GRU-RNN can efficiently capture the long-term dependencies of sequence-based tasks and address the vanishing-gradient problem well [32]. The basic structure of GRU consists of a reset gate r, an update gate z, an activation h, and a candidate activation $\tilde{h}$. Intuitively, the reset gate determines how to combine the new input with the previous memory, and the update gate defines how much of the previous memory to keep around. If we set the reset to all 1's and update gate to all 0's, we arrive at our plain RNN model.

After feeding the sequential LLDs into the GRU-RNNs, we take the last hidden state $\mathbf{h}$ of the top layer as the final representation for classification via a following feedforward and a softmax layer. Owing to the memory capability of GRU-RNNs, the last hidden state of GRU-RNNs can be assumed to be a compressed vector that stores the complete acoustic information over time of a vocalisation [33].

3.4. End-to-End Modelling

Inspired by the work in [34], we further investigated an end-to-end system for our (non-)linguistic classification task. Fig. 1 shows the basic structure of the end-to-end system we employed in this paper. Different from the aforementioned GRU-RNNs modelling, the end-to-end system replaces the inputs from the hand-engineered LLDs with an automatic representation-extraction system. Since the original acoustic signals arguably contain the most complete information, an optimised system can extract the most useful representations for the task of interest, without expert intervention.

Here, the automatic representation extraction system comprised a series of alternating convolutional and max-pooling layers, as shown in Fig. 1. The convolutional layer was uni-dimensional, due to the characteristics of acoustic signals. To train the network, we split the raw waveform into chunks of 40 ms each, and then fed them into the convolutional network.

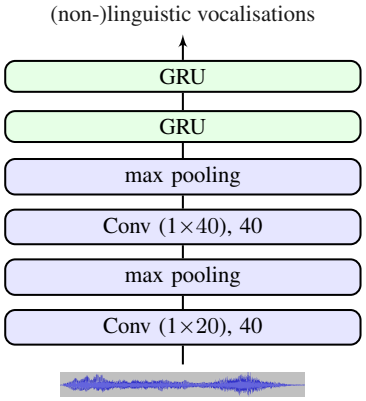


Figure 1: *End-to-end framework for non-/linguistic vocalisation classification.*

The automatically learnt representations were then fed into a joint GRU-RNN similar to the ones described in Section 3.3.

4. Experiments and Results

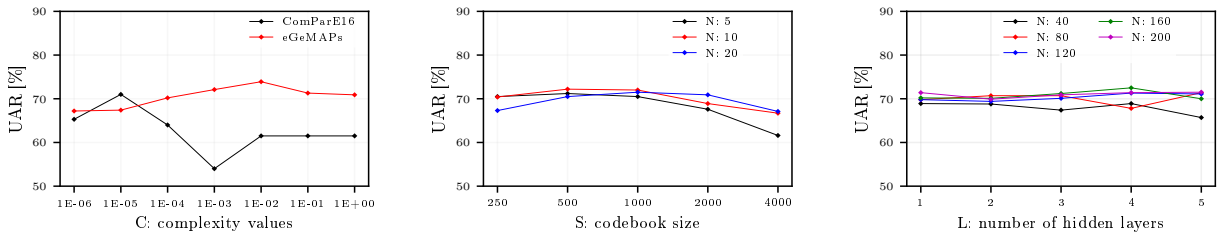
To evaluate the performance of the investigated systems, we utilised the metric of *Unweighted Average Recall* (UAR), which is widely used for computational paralinguistics [28, 26]. The UAR is calculated by the sum of recalls per class divided by the class number, and thus can reflect a meaningful overall accuracy despite class imbalances (such as the one we are facing). It is because we often care more about the model performance over each class rather than the performance on a particular class. Thus, the Weighted Average Recall (WAR, or accuracy) is provided as a complementary metric. Note that, in our experiments we could not compare our systems with the LENA software because the latter can only be used on recordings gathered with the LENA hardware.

4.1. Results of Static Modelling with Functional-based Features

Generally speaking, the eGeMAPS-based and the ComParE16-based functional feature sets perform comparably, with the eGeMAPs achieving the highest result of 73.9% UAR on the development set and the ComParE16 achieving the slightly higher result of 67.7% UAR on the test set (Table 3). This indicates that eGeMAPS indeed contains not only concise but also effective features. Thus, it can be easily learnt by a simple classifier (i. e., SVM in our case). In more detail, when increasing the learning complexity value, the SVM performs increasingly better with the eGeMAPS feature set; however, the opposite is obtained with the ComParE16 feature set (Fig. 2 (a)). This is consistent with our expectation that high-dimensional feature sets normally require lower complexity values but low-dimensional feature sets require higher values [35].

4.2. Results of Static Modelling with Bag-of-Audio-Words

For the experiments, we utilised our open toolkit – openXBOW to extract the BoAW features [29]. The classifier – SVM – was optimised with the complexity value of 10E-4. Figure 2 (b) shows the obtained UARs on the development set by using BoAW representations that were derived from the eGeMAPS LLDs. Both the codebook size (S) and the number of nearest codewords (N) when assigning each audio-word (frame-wise LLDs) are considered, in order to assess the robustness of BoAW representations. It can be seen that the size of the codebook has a positive correlation with the number of nearest assigned codeword when achieving better performance. That is,



(a) Functional-based features + SVM (b) BoAW + SVM (c) LLDs + GRU-RNN

Figure 2: *Performance (UAR) on the development set of (a) the static model (i. e., SVM) with functional-based acoustic feature sets (i. e., eGeMAPS and ComParE16), (b) the static model (i. e., SVM) with BoAW (N: number of nearest codewords assigned for each audio-word), and (c) the dynamic model (i. e., GRU-RNN) with sequential LLDs (N: number of nodes per hidden layer).*

the smaller codebook (e. g., 250) requires a smaller number of assigned codewords (i. e., 5/10 in our case); whereas the larger codebook (e. g., 2 000) needs to assign each audio-word to more nearest codewords (i. e., 20 in our case).

4.3. Results of Dynamic GRU-RNN Modelling

For training the networks, we employed the Adam optimisation algorithm with an optimised learning rate of 10E-4. The batch size was set to 128 to facilitate the training process. Figure 2 (c) shows the performance of GRU-RNN systems on the development set with a set of hidden layers and nodes per hidden layer. Again, the eGeMAPS LLDs were employed as the network inputs. Generally speaking, the GRU-RNN modelling performs stably when alternating the network structures. The best performance (i. e., 72.5 %) was yielded by the network structure of 4 hidden layers and 160 nodes per layer.

4.4. Results of End-to-End Modelling

We conducted these experiments using our open toolbox – End2You [34, 36]. The default network was optimised for emotion recognition, which consists of two convolutional layers (each of them followed by a max pooling layer) and two GRU recurrent layers. Specifically, all audio files were resampled into 16 kHz. The first convolutional layer used 20 space time finite impulse filters with a 5 ms window in order to extract the fine-scale information, and the second convolutional layer used 40 space time finite impulse filters with a 500 ms window to extract a long-term characteristics of speech. Both convolutional layers included 40 filters. Additionally, the pool size of the first and the second max pooling layers were set to be 2 and 10 to reduce the sampling rate.

4.5. Discussion

The best results (UARs and WARs) on the development and test sets for each system are shown in Table 3, which also includes a simple baseline of the type a trained phonetician may generate.[2] Generally speaking, all investigated systems significantly outperform this simple baseline system (one-sided z-test, all $p < .05$). The highest UAR on the test set 70.4 % was achieved with dynamic GRU-RNN model on ComParE16 LLDs.

As for the eGeMAPS versus ComParE16 LLDs, the former perform better on the development set, while the opposite is true for the test set. The small performance gap between the development set and test set implicitly indicates that the ComParE16 feature set (including much more acoustic information) helps the system become more robust and stable. The static SVM models and the dynamic GRN-RNN model perform similarly. Moreover, we found that the end-to-end system did not outperform the other systems. Future work could change this by using more training data, and optimising the network structure.

[2]The second author assumes full responsibility for this baseline. Using Praat [37], formant estimations were drawn by fitting two poles below 4 kHz. The rule generalised from the dev set was: if less than a third of the frames had missing formant data, the clip was labelled as "linguistic", otherwise as "non-linguistic".

Table 3: *Performance comparison among the investigated systems against a simple baseline.*

[%]	dev		test	
approaches (LLDs)	UAR	WAR	UAR	WAR
Baseline (missing formants)	44.2	61.9	43.8	56.6
Functionals (eGeMAPS)	**73.9**	79.9	66.8	74.0
Functionals (ComParE16)	71.0	79.8	67.7	76.8
BoAW (eGeMAPS)	72.2	70.8	68.4	67.1
BoAW (ComParE16)	71.3	71.4	69.5	71.2
GRU-RNN (eGeMAPS)	72.5	74.3	66.7	70.6
GRU-RNN (ComParE16)	71.3	71.0	**70.4**	70.5
End-to-end	61.9	**83.6**	63.2	**80.1**

5. Conclusions

To identify the linguistic or non-linguistic vocalisations from the daylong recordings, there is a research trend shifting from a laborious manual-counting strategy to automated classification. In this paper, we investigated several different systems with a variety of feature types. From the experimental results, we found that most of the employed systems can achieve promising recognition accuracies. We further found that the ComParE16 acoustic feature set is more helpful for creating a robust and stable system than eGeMAPS. Moreover, we found that the available end-to-end system has the potential to capture the effective features from scratch, but does not yet outperform systems trained with hand-engineered features. In future work, we intend to extend these approaches to a finer-grained classification of this dataset, which may allow a closer connection with the psycholinguistic literature on early vocal development.

6. Acknowledgements

This work was supported by a TransAtlantic Platform "Digging into Data" collaboration grant (ACLEW: Analyzing Child Language Experiences Around The World), with the support of the UK's Economic & Social Research Council through the research Grant No. HJ-253479 (ACLEW); the French Agence Nationale de la Recherche (ANR-16-DATA-0004 ACLEW; ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC); the National Science Foundation (BCS-1529127 and SMA-1539129); and the James S. McDonnell Foundation.

7. References

[1] D. K. Oller, *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[2] D. K. Oller, E. H. Buder, H. L. Ramsdell, A. S. Warlaumont, L. Chorna, and R. Bakeman, "Functional flexibility of infant vocalization and the emergence of language," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 16, pp. 6318–6323, 2013.

[3] R. E. Stark, "Stages of speech development in the first year of life," in *Child Phonology, Vol. 1: Production*, G. H. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson, Eds. New York, NY: Academic Press, 1980, pp. 73–92.

[4] D. K. Oller, "The emergence of the sounds of speech in infancy," in *Child Phonology, Vol. 1: Production*, G. H. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson, Eds. New York, NY: Academic Press, 1980, pp. 93–112.

[5] L. Roug, I. Landberg, and L.-J. Lundberg, "Phonetic development in early infancy: A study of four Swedish children during the first eighteen months of life," *Journal of Child Language*, vol. 16, no. 1, pp. 19–40, 1989.

[6] F. J. Koomans-van Beinum and J. M. van der Stelt, "Early stages in the development of speech movements," in *Precursors of early speech*, B. Lindblom and R. Zetterström, Eds. New York, NY: Stockton Press, 1986, pp. 37–50.

[7] A. S. Warlaumont, J. A. Richards, J. Gilkerson, and D. K. Oller, "A social feedback loop for speech development and its reduction in autism," *Psychological Science*, vol. 25, no. 7, pp. 1314–1324, 2014.

[8] Y. Shinya, M. Kawai, F. Niwa, and M. Myowa-Yamakoshi, "Preterm birth is associated with an increased fundamental frequency of spontaneous crying in human infants at term-equivalent age," *Biology Letters*, vol. 10, no. 8, p. 20140350, 2014.

[9] P. F. MacNeilage, "The frame/content theory of evolution of speech production," *Behavioral and Brain Sciences*, vol. 21, no. 4, pp. 499–511, Aug. 1998.

[10] G. A. Bryant and C. A. Aktipis, "The animal nature of spontaneous human laughter," *Evolution and Human Behavior*, vol. 35, no. 4, pp. 327–335, July 2014.

[11] A. S. Warlaumont and M. F. Finnegan, "Learning to produce syllabic speech sounds via reward-modulated neural plasticity," *PLOS ONE*, vol. 11, no. 1, p. e0145096, Jan. 2016.

[12] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. Hansen, "Signal processing for young child speech language development," in *Proc. 1st Workshop on Child, Computer, and Interaction*, Chania, Greece, 2008.

[13] E. Scheiner, K. Hammerschmidt, U. Jürgens, and P. Zwirner, "The influence of hearing impairment on preverbal emotional vocalizations of infants," *Folia phoniatrica et logopaedica*, vol. 56, no. 1, pp. 27–40, Feb. 2004.

[14] C. Z. Malatesta, "Infant emotion and the vocal affect lexicon," *Motivation and Emotion*, vol. 5, no. 1, pp. 1–23, Mar. 1981.

[15] L. J. Trainor, C. M. Austin, and R. N. Desjardins, "Is infant-directed speech prosody a result of the vocal expression of emotion?" *Psychological Science*, vol. 11, no. 3, pp. 188–195, May 2000.

[16] P. M. Cole, M. K. Michel, and L. O. Teti, "The development of emotion regulation and dysregulation: A clinical perspective," *Monographs of the society for research in child development*, vol. 59, no. 2-3, pp. 73–102, Feb. 1994.

[17] P. Pal, A. N. Iyer, and R. E. Yantorno, "Emotion detection from infant facial expressions and cries," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. II–721–II–724.

[18] H. J. Fell, J. MacAuslan, L. J. Ferrier, S. G. Worst, and K. Chenausky, "Vocalization age as a clinical tool," in *Proc. 7th International Conference on Spoken Language Processing (ICSLP '02)*, Denver, Colorado, 2002, pp. 2345–2348.

[19] A. S. Warlaumont, D. K. Oller, E. H. Buder, R. Dale, and R. Kozma, "Data-driven automated acoustic analysis of human infant vocalizations using neural network tools," *The Journal of the Acoustical Society of America*, vol. 127, p. 2563, 2010.

[20] A. S. Warlaumont and H. L. Ramsdell-Hudock, "Detection of total syllables and canonical syllables in infant vocalizations." in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 2676–2680.

[21] G. Yetish, J. Siegel, and A. Cristia, "Daylong recordings from young children learning Ju'hoan in Namibia," https://nyu.databrary.org/volume/446, accessed: 2018-03-03.

[22] C. Scaff, J. Stieglitz, and A. Cristia, "Daylong recordings from young children learning Tsimane in Bolivia," https://nyu.databrary.org/volume/445, accessed: 2018-03-03.

[23] H. Colleran and A. Cristia, "Daylong recordings from young children learning multiple languages in Vanuatu," https://nyu.databrary.org/volume/449, accessed: 2018-03-03.

[24] Z. Zhang, J. Geiger, J. Pohjalainen, E. D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Transactions on Intelligent Systems and Technology*, 2018, 23 pages, to appear.

[25] B. Schuller, "The computational paralinguistics challenge," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 97–101, July 2012.

[26] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. C. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language." in *Proc. INTERSPEECH*, Portland, OR, 2016, pp. 2001–2005.

[27] F. Eyben, K. R. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

[28] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, United Kingdom, 2009, pp. 312–315.

[29] M. Schmitt and B. Schuller, "openXBOW – introducing the Passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, no. 96, pp. 1–5, Oct. 2017.

[30] J. Han, Z. Zhang, Z. Ren, F. Ringeval, and B. Schuller, "Bags in bag: Generating context-aware bags for tracking emotions from speech," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, 5 pages.

[31] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, Doha, Qatar, 2014, pp. 103–111.

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, Dec. 2014.

[33] Z. Zhang, D. Liu, J. Han, and B. Schuller, "Learning audio sequence representations for acoustic event classification," *arXiv preprint arXiv:1707.08729*, July 2017.

[34] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[35] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer-Verlag, 1995.

[36] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, "End2You – The Imperial toolkit for multimodal profiling by end-to-end learning," *arXiv preprint arXiv:1802.01115*, Feb. 2018.

[37] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat.org/*, 2006.