VCMNet: Weakly Supervised Learning for Automatic Classification of Infant Vocalisations

Najla D. Al Futaisi
GLAM – Group on Language, Audio
Music, Imperial College London
London, UK
n.al-futaisi18@imperial.ac.uk

Zixing Zhang
GLAM – Group on Language, Audio
Music, Imperial College London
London, UK
zixing.zhang@imperial.ac.uk

Alejandrina Cristia
LSCP, Département d'études
cognitives, ENS, EHESS, CNRS, PSL
Research University
Paris, France
alecristia@gmail.com

Anne S. Warlaumont
Department of Communication
University of California, Los Angeles
Los Angeles, USA
warlaumont@ucla.edu

Björn W. Schuller
GLAM – Group on Language, Audio
Music, Imperial College London
London, UK
University of Augsburg
Augsburg, Germany
bjoern.schuller@imperial.ac.uk

ABSTRACT

Using neural networks to classify infant vocalisations into important subclasses (such as crying versus speech) is an emergent task in speech technology. One of the biggest roadblocks standing in the way of progress lies in the datasets: The performance of a learning model is affected by the labelling quality and size of the dataset used, and infant vocalisation datasets with good quality labels tend to be small. In this paper, we assess the performance of three models for infant VoCalisations Maturity (VCM) trained with a large dataset annotated automatically using a purpose-built classifier and a small dataset annotated by highly trained human coders. The two datasets are used in three different training strategies, whose performance is compared against a baseline model. The first training strategy investigates adversarial training, while the second exploits multi-task learning as the neural network trains on both datasets simultaneously. In the final strategy, we integrate adversarial training and multi-task learning. All of the training strategies outperform the baseline, with the adversarial training strategy yielding the best results on the development set.

KEYWORDS

infant vocalisation, prelinguistic analysis, weakly supervised learning

ACM Reference Format:
Najla D. Al Futaisi, Zixing Zhang, Alejandrina Cristia, Anne S. Warlaumont, and Björn W. Schuller. 2019. VCMNet: Weakly Supervised Learning for Automatic Classification of Infant Vocalisations. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340555.3353751

1 INTRODUCTION

From birth, infants start producing vocalisations, some of which are not linguistic – notably crying. Among those that are linguistic, canonical babbles which typically emerge before 10 months of age, consist of syllables with a vowel and a consonant sound [11] such as 'dada' are more complex than non-canonical ones like 'aaah'. In broad terms, the proportion of vocalisations that are linguistic, as well as the proportion of linguistic vocalisations that are canonical as opposed to non-canonical, both increase with age [12, 19, 20]. Studying *infant VoCalisation Maturity* (VCM) is crucial for early detection of language impairment risks, and to describe potential group differences such as those that may appear between children with/without a family history of impairment. Such research is also relevant from a theoretical standpoint: If children who exhibit abnormal early vocal development have delayed or impaired language later on, this supports the theory that early vocal patterns lay the foundation of later language [10]. Recent research has exploited day-long audio recordings gathered with an infant-worn device because these can capture many samples of the child's natural production [6]; however, in most cases these 8-16h long recordings cannot feasibly be completely annotated manually. Therefore, a current goal is to improve automatic segmentation and classification procedures [24].

One of the major problems for VCM classification is the lack of carefully annotated data (see [19, 24] and references therein for small-scale approaches). This is particularly troublesome for deep learning approaches, which normally require large-scale data to extract robust representations. One potential solution is to leverage massive unlabelled or weakly supervised data. Here we propose a weakly supervised framework, namely *Adversarial Multi-Task Learning* (AMTL) to relax the high requirement of burdensome human-annotated data (*aka* strongly labelled data), by exploiting massive data (*aka* weakly labelled data) automatically annotated by a third-party system, i. e., the LENA system [22]. The LENA system

ICMI '19, October 14–18, 2019, Suzhou, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3353751

segmented and classified the audio according to sound source type; these data are in many cases inaccurately segmented and classified, and furthermore use different VCM categories. The proposed AMTL framework combines two main learning strategies, i. e., the multi-task learning (MTL) and the domain adversarial training. The first tries to transfer the knowledge from the weakly labelled data to our target task, whereas the latter strategy aims to alleviate the domain mismatch between the two data distributions. Despite the fact that AMTL has been introduced for text classification [8], this is the first time to be investigated in the voice domain, to the best of our knowledge.

2 METHODOLOGY

2.1 Multi-Task Learning (MTL)

The multi-task learning strategy jointly trains the model on several different but relevant tasks simultaneously. In our case, the two tasks refer to (1) a four-class *target* task on the strongly human-labelled dataset (see Section 3.1 for more details); and (2) a two-class *source* task on the weakly-labelled dataset (see Section 3.2 for more details). MTL is an effective approach to distill robust representations shared across various tasks, and knowledge can be potentially transferred from other source tasks to the target one [23]. This is particularly beneficial to the target task when it has limited training samples, because MTL strategy is able to partially overcome the overfitting problem [25]. The middle and right paths in Figure 1 show the network structure for the four-class VCM task (target task) and the two-class VCM task (source task), respectively.

Mathematically, the loss function of MTL can be written as the following equation:

$$\mathcal{J}(\theta_g, \theta_s, \theta_w) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_s^i(\theta_g, \theta_s) + \alpha\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_w^i(\theta_g, \theta_w), \tag{1}$$

where n and m denotes the number of strongly labelled data and weakly labelled data; θ_g, θ_s, θ_w, respectively, represent the network parameters of the feature extraction layers, of the classification layers for the four-class target task trained on the strongly labelled data, and of the classification layers for the two-class source task trained on the weakly labelled data. $\mathcal{L}_s^i$ and $\mathcal{L}_w^i$ stand for the losses (i. e., cross entropy) from target and source tasks, respectively, while the hyper-parameter α controls the contributions from the source task.

2.2 Adversarial Training

Despite the fact that MTL can transfer knowledge between multiple tasks, it depends on the assumption that the tasks are from the same or similar domains [8]. That is, it cannot guarantee that the learnt high-level representations share the same space if the inputs come from different domains [7, 8]. In this case, the underlying information cannot freely flow over different tasks. To address the domain mismatch issue, domain adversarial training, proposed by Ganin et al. in 2014 [5], aims to learn the domain invariant features among different data domains. In our cases, there are the strongly labelled dataset (source domain) and the weakly labelled dataset (target domain).

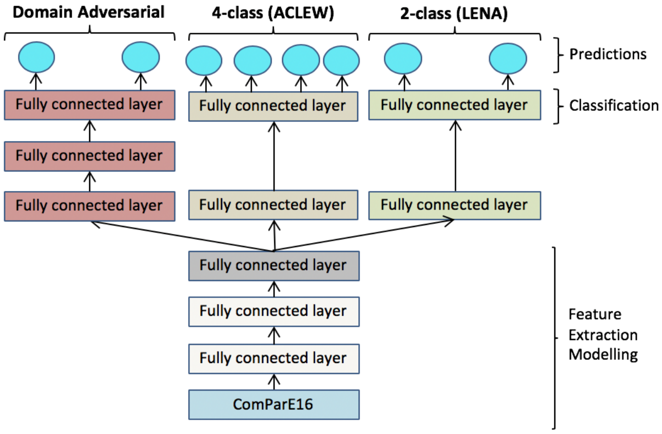

Figure 1: Framework of the introduced adversarial multi-task learning for infant vocal maturity analysis

The architecture of the domain adversarial training is displayed in Figure 1 with the middle and left paths. The feature extractor projects the data from different separate domains into high-level representations, which are discriminative for a VCM classifier (middle path) and indistinguishable for a domain classifier (left path). The labels for the domain adversarial task correspond to the domains, i.e. the datasets. We arbitrarily defined the labels of strongly labelled data to be 0, and the labels of weakly labelled data to be 1. Mathematically, the networks are optimised by the following objective function:

$$\begin{aligned}\mathcal{J}(\theta_g, \theta_s, \theta_d) =& \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_s^i(\theta_g, \theta_s) - \\ & \beta(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_d^i(\theta_g, \theta_d) + \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_d^i(\theta_g, \theta_d)),\end{aligned} \tag{2}$$

where $\mathcal{L}_s$ and $\mathcal{L}_d$ denote the classification loss from the VCM classifier and the domain classifier. A hyper-parameter *beta* is utilised to tune the trade-off between the two tasks during the learning process. Particularly, the parameters of the feature extractor (θ_g) and VCM classifier (θ_s) are obtained by minimising the objective function as

$$(\hat{\theta}_g, \hat{\theta}_s) = \arg\min_{\theta} \mathcal{J}(\theta_g, \theta_s, \hat{\theta}_d) \tag{3}$$

In contrast, the parameters of the domain classifier (θ_d) are trained by maximising the objective function as

$$\hat{\theta}_d = \arg\max_{\theta} \mathcal{J}(\hat{\theta}_g, \hat{\theta}_s, \theta_d) \tag{4}$$

With such a minmax optimisation process, the representations learnt from different domains cannot be easily distinguished [5]. Due to its efficiency, domain adversarial training is being used widely in diverse applications, such as affective computing [7].

2.3 Adversarial Multi-Task Learning (AMTL)

As aforementioned, domain adversarial training is able to distil the domain invariant representations, whereas MTL is efficient in extracting robust and discriminative representation among different tasks. Adversarial multi-task learning takes advantages of both learning strategies. Figure 1 shows the complete network structure of AMTL.

Overall, the network parameters except the ones of the domain classifiers are optimised by minimising this objective function:

$$\mathcal{J}(\theta_g,\theta_s,\theta_d) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_s^i(\theta_g,\theta_s) + \alpha\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_w^i(\theta_g,\theta_w) - \beta(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_d^i(\theta_g,\theta_d) + \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_d^i(\theta_g,\theta_d)) \qquad (5)$$

Again, the hyper-parameters of α and β control the contributions from MTL and domain adversarial training, respectively.

3 DATASETS

The next two subsections provide more detailed descriptions of the two datasets used. They were partitioned into training, development and test sets as shown in Table 1.

3.1 Strongly Labelled Dataset

The strongly labelled dataset is human labelled. It is the union of several datasets. One group was collected using infant-worn recorders that gathered audio data for whole days from 60 children aged 3-36 months. Half heard US/Canadian English ([1, 9, 18]; some in this group additionally heard Spanish or French); 10 children were exposed to UK English [15], 10 to Tseltal [2], 10 to Argentinean Spanish [14]; some of these day-long recordings are available from HomeBank [17]. For each child, 15 2-minute long clips were randomly sampled, segmented, and annotated using the ACLEW annotation scheme [3], which includes two linguistic classes (canonical and noncanonical vocalisations) and some non-linguistic classes (crying, laughing). The rest of the data was collected in the lab from 15 English-learning infants aged 8-16 months [13]. These datasets are merged into one, to increase the size of the dataset and obtain better-performing models, which we refer to as the strongly human labelled ACLEW dataset.

3.2 Weakly Labelled Dataset

The weakly labelled dataset came from the 20 day-long US recordings [1, 18], labelled in their entirety using the LENA system [22]. This system was developed to analyse day-long infant-centred recordings, and includes a two-class linguistic versus non-linguistic classification [21]. Linguistic vocalisations include canonical and non-canonical; non-linguistic ones include crying, laughing, and vegetative sounds such as burping. Previous work [21] found that the precision of the LENA algorithm for (non-)linguistic was 75% and 84%, respectively. We refer to this dataset as the weakly labelled LENA dataset.

4 EXPERIMENTS AND RESULTS

In this section, we conducted tentative experiments on the datasets as described in Section 3 to evaluate the performance of the approaches introduced in Section 2.

Table 1: Data distribution over different partitions and categories of the ACLEW and LENA datasets. Non-can stands for non-canonical; can for canonical.

ACLEW	∑	**non-can.**	**can.**	**crying**	**other**
Train	8 194	5 664	2 156	263	112
Develop	4 573	3 076	1 250	210	37
Test	4 060	2 956	827	234	44
∑	16 827	11 696	4 233	707	193
LENA	∑	**ling.** **(non-can./can.)**		**non-ling.** **(crying/other)**	
Train	28 572	17 012		11 560	
Develop	4 317	2 939		1 378	
Test	5 017	3 123		1 894	
∑	37 906	23 074		14 832	

4.1 Experimental Setups

Based on the onset and offset information of VCM annotations, we segmented each sample from the day-long recordings, leading to 16 827 and 37 906 samples of strongly labelled data (ACLEW) and weakly labelled data (LENA), respectively. We randomly (without considering which participant a sample was drawn from) split these samples into training, development, and test sets. The data distributions over the different partitions and VCM categories of the ACLEW and LENA datasets are shown in Table 1.

We applied the widely used openSMILE [4] toolkit to extract acoustic features, specifically the feature set designed for the Interspeech 2016 Computational Paralinguistics Challenge (ComParE16) [16]. To obtain these features, we firstly extracted 65 low-level descriptors (LLDs), as well as their first derivation (delta) at the frame level, resulting in 130 LLDs per frame. Then, we applied a set of functionals such as extremes, means, moments, percentiles, and peaks to the sequential LLDs, resulting in 6 373 dimensional static features per segment.

To evaluate the performance, we employed the frequently used metrics of macro F1 and Unweighted Average Recall (UAR) for VCM classification. Both UAR and F1 are calculated by the sum of classwise recall or F1 divided by the number of classes. Thus, it indicates the system performance in an unbalanced data distribution case.

4.2 Network Training

When designing the framework, we used the conventional deep neural network structure, which consists of multiple feed-forward fully connected layers (dense layers) consisting of 100 nodes each, with a random weight initialisation. ReLU activation function was employed to address the overfitting and gradient vanishing problems. An L2 norm regularisation term with a control weight of 0.0001 was added to punish the weights with high values and a dropout rate of 0.2 was applied to each dense layer, so as to further reduce overfitting. We used three hidden layers for the shared network (feature extracting model), and two hidden layers for each of the classification tasks. When training the network, Adam optimiser was selected with an initial learning rate of 0.001. A batch size of 128 per training iteration was used to facilitate efficient learning. All these hyper-parameters were determined by a random search strategy for the best performing model on the development set of the ACLEW dataset.

Before feeding the training data into the network, an online standardisation strategy was employed, by applying the mean and variance obtained from the ACLEW training set to the development and test sets. For the LENA dataset, only the training set was used in our experiments and it was standardised by using the LENA training dataset mean and variance.

For AMTL, we train all three branches simultaneously, as each training batch includes both weakly and strongly labelled data, using Eq. 5 to optimise the whole network. The gradient reversal operation is employed when merging all the losses to optimise the feature extractor layer, as shown in Eq. 5.

4.3 Results and Discussions

Results are shown in Table 2. Note that all evaluations are based on the results for the development and test sets of the ACLEW dataset. The baseline system refers to the implementation without any MTL or domain adversarial training processes. Basically, it can be seen that the results obtained with the baseline system are much higher than the chance level (i. e., 25 % of F1 or UAR), which indicates that automatically assessing the maturity of infant vocalisation can be fairly easily done at an above-chance level. Nevertheless, there is still much room for improvement.

With the MTL strategy, the results are higher, and the UARs in particular are significantly higher (two-tailed z-test, $p < .05$), than with the baseline system; the UARs have been boosted from 47.7 % and 45.5 % to 54.3 % and 50.1 % on the development and test sets, respectively. This implies that some knowledge is distilled from the weakly labelled dataset (LENA) dataset to the strongly labelled (ACLEW) dataset via the extraction of shared acoustic representations, despite the former dataset having a different and less accurate annotation scheme.

The domain adversarial training approach also significantly outperforms (two-tailed z-test, $p < .05$) the baseline system in three out of four cases. The benefit might stem from the fact that the large-scale auxiliary weakly labelled dataset extends the representation space extracted from a small dataset, while reducing the domain mismatch among different training sets, avoiding overfitting to the original training set to a certain degree.

Finally, when integrating the MTL approach and the domain adversarial training approach, one can see that the proposed AMTL system was significantly (two-tailed z-test, $p < .05$) superior to the baseline system in all scenarios. Nevertheless, AMTL did not always yield the best results among all investigated approaches. This might partially relate to the fact that a lot of data are wrongly segmented or annotated by the LENA system, limiting the effectiveness of MTL.

Table 2: Performance comparison in terms of F1 and UAR among the learning strategies. The cases where the investigated systems have a statistical significance of performance improvement over the baseline system via a two-tailed z-test are marked by the "$*$" symbol. Bold indicates best performance for each metric and set.

[%]	develop		test	
approaches	F1	UAR	F1	UAR
Baseline	49.5	47.7	47.4	45.5
Multi-task learning (MTL)	50.9	54.3*	48.6	**50.1***
Domain adversarial train.	**51.9***	**54.6***	49.5	50.0*
Adversarial MTL	51.7*	53.9*	**49.6***	50.0*

5 CONCLUSIONS

To overcome the data sparsity problem for analysing the maturity of infant vocalisations, we proposed an adversarial multi-task learning approach to leverage the value of numerous weakly labelled data, automatically annotated by the commercial LENA system [22]. Despite the messy characteristic of these augmented data, we obtained a significant performance improvement compared with baseline systems. In the future we aim to include a preprocessing step to remove the wrongly-segmented or -labelled data prior to training. Moreover, we may also consider integrating this approach with other weakly supervised learning structures, e.g. auto-encoders.

6 ACKNOWLEDGEMENTS

This work was supported by a Trans-Atlantic 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-1539129/1827744); and the James S. McDonnell Foundation.

REFERENCES

[1] Elika Bergelson. 2016. Bergelson Seedlings HomeBank Corpus. https://doi.org/10.21415/T5PK6D

[2] Marisa Casillas, Penelope Brown, and Steven C. Levinson. 2017. Casillas HomeBank Corpus. https://homebank.talkbank.org/access/Secure/Casillas.html

[3] Marisa Casillas, John Bunce, Melanie Soderstrom, Celia Rosemberg, Maia Migdalek, Florencia Alam, Alejandra Stein, and Hallie Garrison. 2017. Introduction: The ACLEW DAS template. https://osf.io/aknjv/

[4] Florian Eyben, Martin Wöllmer, and Björn W. Schuller. 2010. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, Florence, Italy, 1459–1462.

[5] Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495* (2014).

[6] Jill Gilkerson and Jeffrey A Richards. 2008. The LENA natural language study. *Boulder, CO: LENA Foundation. Retrieved March* 3 (2008), 2009.

[7] Jing Han, Zixing Zhang, Nicholas Cummins, and Björn W. Schuller. 2018. Adversarial Training in Affective Computing and Sentiment Analysis: Recent Advances and Perspectives. *arXiv preprint arXiv:1809.08927* (2018).

[8] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742* (2017).

[9] Karmen McDivitt and Melanie Soderstrom. 2016. McDivitt HomeBank Corpus. https://doi.org/10.21415/T5KK6G

[10] D Kimbrough Oller. 2000. *The emergence of the speech capacity*. Psychology Press.

[11] D Kimbrough Oller, Rebecca E Eilers, A Rebecca Neal, and Alan B Cobo-Lewis. 1998. Late onset canonical babbling: A possible early marker of abnormal development. *American Journal on Mental Retardation* 103, 3 (1998), 249–263.

[12] D Kimbrough Oller, Rebecca E Eilers, Richard Urbano, and Alan B Cobo-Lewis. 1997. Development of precursors to speech in infants exposed to two languages. *Journal of child language* 24, 2 (1997), 407–425.

[13] Heather L. Ramsdell-Hudock, Andrew Stuart, and Douglas F. Parham. 2018. Utterance Duration as It Relates to Communicative Variables in Infant Vocal Development. *Journal of Speech, Language, and Hearing Research* 61, 2 (2018), 246–256.

[14] Celia R. Rosemberg, Florencia Alam, Alejandra Stein, Migdalek Maia., Alejandra Menti, and Gladys Ojea. 2015. Los entornos lingüísticos de niñas y niños pequeños argentinos / Language Environments of Young Argentinean Children.

[15] Caroline F. Rowland, Amy Bidgood, Samantha Durrant, Michelle Peter, and Julian M Pine. 2018. The Language 0-5 Project. https://doi.org/10.17605/OSF.IO/KAU5F

[16] Björn W. Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron C. Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. In *INTERSPEECH*. 2001–2005.

[17] Mark VanDam, Anne S. Warlaumont, Eric Bergelson, Alexandrina Cristia, Melanie Soderstrom, Paul D. Palma, and Brian MacWhinney. 2016. HomeBank: An online repository of daylong child-centered audio recordings. *Seminars in Speech and Language* 37, 128–142. Issue 2.

[18] Anne S. Warlaumont, Gine M Pretzer, Sara Mendoza, and Eric A. Walle. 2016. Warlaumont HomeBank Corpus. https://doi.org/10.21415/T54S3C

[19] Anne S. Warlaumont and Heather L. Ramsdell-Hudock. 2016. Detection of Total Syllables and Canonical Syllables in Infant Vocalizations. In *INTERSPEECH*. 2676–2680.

[20] Anne S. Warlaumont, Jeffrey A. Richards, Jill Gilkerson, and D. Kimbrough Oller. 2014. A social feedback loop for speech development and its reduction in autism. *Psychological science* 25, 7 (2014), 1314–1324.

[21] Dongxin Xu, Umit Yapanel, and Sharmi Gray. 2009. Reliability of the LENA Language Environment Analysis System in young children's natural home environment. *Boulder, CO: LENA Foundation* (2009), 1–16.

[22] Dongxin Xu, Umit Yapanel, Sharmi Gray, Jill Gilkerson, Jeff Richards, and John Hansen. 2008. Signal processing for young child speech language development. In *First Workshop on Child, Computer and Interaction*.

[23] Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114* (2017).

[24] Zixing Zhang, Alejandrina Cristia, Anne S. Warlaumont, and Björn W. Schuller. 2018. Automated Classification of Children's Linguistic versus Non-Linguistic Vocalisations. In *Proc. Interspeech 2018*. 2588–2592. https://doi.org/10.21437/Interspeech.2018-2523

[25] Zixing Zhang, Bingwen Wu, and Björn W. Schuller. 2019. Attention-augmented End-to-end Multi-task Learning for Emotion Prediction from Speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6705–6709.