Using Attention Networks and Adversarial Augmentation for Styrian Dialect Continuous Sleepiness and Baby Sound Recognition

Sung-Lin Yeh[1,2], *Gao-Yi Chao*[1,2], *Bo-Hao Su*[1,2], *Yu-Lin Huang*[1,2], *Meng-Han Lin*[1,2], *Yin-Chun Tsai*[1,2], *Yu-Wen Tai*[1], *Zheng-Chi Lu*[1], *Chieh-Yu Chen*[3], *Tsung-Ming Tai*[3], *Chiu-Wang Tseng*[3], *Cheng-Kuang Lee*[3], *Chi-Chun Lee*[1,2]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
[3]NVIDIA, Taiwan

cclee@ee.nthu.edu.tw

Abstract

In this study, we present extensive attention-based networks with data augmentation methods to participate in the INTERSPEECH 2019 ComParE Challenge, specifically the three Sub-challenges: Styrian Dialect Recognition, Continuous Sleepiness Regression, and Baby Sound Classification. For Styrian Dialect Sub-challenge, these dialects are classified into Northern Styrian (NorthernS), Urban Sytrian (UrbanS), and Eastern Styrian (EasternS). Our proposed model achieves an UAR 49.5% on the test set, which is 2.5% higher than the baseline. For Continuous Sleepiness Sub-challenge, it is defined as a regression task with score range from 1 (extremely alert) to 9 (very sleepy). In this work, our proposed architecture achieves a Spearman correlation 0.369 on the test set, which surpasses the baseline model by 0.026. For Baby Sound Sub-challenge, the infant sounds are classified into canonical babbling, non-canonical babbling, crying, laughing and junk/other, and our proposed augmentation framework achieves an UAR of 62.27% on the test set, which outperforms the baseline by about 3.5%. Overall, our analyses demonstrate that by fusing attention network models with conventional support vector machine benefits the test set robustness, and the recognition rates of these paralinguistic attributes generally improve when performing data augmentation.

Index Terms: attention networks, augmentation, adversarial learning, computational paralinguistics

1. Introduction

In recent years, development of speech acoustic algorithms has helped advance many human-machine interface designs across applications such as companion robot [1, 2], meeting assistant [3, 4], and medical agent [5]. Furthermore, the ability to computationally extract paralinguistic information conveyed in these recorded acoustic signals is important in designing novel engineering solutions that were not possible before. In fact, ComParE Challenges have been committed to provide a platform for a wealth paralinguistic attributes recognition benchmarks, e.g., emotion, gender, age, autism spectrum disorder, etc., using real-world dataset in the past decade. In this year, ComParE Challenge 2019 is composed of four sub-challenges: Styrian Dialect Sub-Challenge, Continuous Sleepiness Sub-challenge, Baby Sound sub-challenge, and Orca Activity Sub-challenge [6]. In this work, we propose three specific algorithmic approaches to participate in Styrian Dialect, Continuous Sleepiness, and Baby Sound Sub-challenges.

Firstly, for Styrian Dialect recognition Sub-challenge, it is well-known that in the same language, the accents and the pronunciations could be different across regions resulting in Dialectal variation. This dialect variation causes an issue in speech recognition and speech or linguistic-based therapeutic assessment [7, 8, 9]. The ability to automatically recognize different dialects is an important technological module in speech-based solutions. Secondly, for Continuous Sleepiness Sub-challenge, insomnia has become a major issue in the modern society, and its severity could even be a prognostic factor for some illnesses [10]. Recognizing the sleepiness level from speech acoustic could help decide the proper therapy and adjust one's life style as early as possible. Lastly, for Baby Sound Sub-challenge, without language ability, making sound the only way that infants could express their feelings. The subtle variation in the acoustics of the baby sound could indicate their emotional and physiological conditions [11, 12]. Developing algorithms that could reliably classify the types of baby sound categorizations can help improve the quality of infant care.

These computational paralingistic tasks often suffer from varying recording conditions, e.g., background noise, recording devices, environment factors, and mismatch and unbalanced class distributions between training and testing. These challenging conditions resemble closely to the real-world application scenarios. In this work, we apply a variety of attention-based networks augmented with generated or real samples for each of the three Sub-challenges. For Styrian Dialect Sub-challenge, we perform decision score fusion between three models: the baseline model, a support vector machine (SVM) trained with volume augmented dataset, and attention-based convolutional neural network (CNN) with volume augmentation as well. Our proposed model achieves 51.7% and 49.5% in development and test set, which improves 7.3% and 2.5% over the baseline model. For Continuous Sleepiness Sub-challenge, we perform decision score fusion between four models: the baseline model, a support vector regressor trained on the extreme subset, attention-based bi-directional long short-term memory (BLSTM), and attention-based CNN. Our proposed framework achieves a Spearman correlation of 0.373 and 0.369, which is an absolute improvement of 0.099 and 0.026 over the baseline model. For Baby Sound Sub-challenge, we use an adversarial auto-encoder network (AAE) to generate samples and additionally include real samples of infant crying to perform training data augmentation. We extract all three types (ComPareE, auDeep, and IS10-paraling) of feature sets to train a SVM model. This approach achieves 61.37% and 62.27% UAR in development and test set, that is a 6.37% and 3.57% absolute improvement over the baseline model.

2. Research Methodology

2.1. Frameworks

2.1.1. Attention-based CNN

CNN-based models have been applied successfully to tasks of speech recognition [13], speech emotion recognition [14, 15], and dialect recognition [16]. In this work, we use eGemaps-LLD as input features with dimension of $(L, D, 1)$ for our CNN model, where L is the sequence length, D is the feature dimension. Each feature can then be treated as a single image with one channel. These input features are passed through 3 convolutional layers to generate feature maps with dimension of (L', D', C'). Feature maps are reshaped into a 2D matrix with dimension $(C', L' \times D')$ then pass to a fully-connected layer. The output of the fully-connected layer M with dimension (C', d) is then treated as the input of the attention layer, where d is the hidden dimension. Then, we perform mean-pooling to the learned attention-weighted feature maps to obtain the hidden representations z, where $z \in \mathbb{R}^d$. z are finally passed to subsequent two dense layers for recognition. In the Styrian Dialect Sub-challenge, the output dimension is three, corresponding to probabilities of the three classes after softmax.

$$Y_{SD} = softmax\left(relu\left(zW_1 + b_1\right)W_2 + b_2\right) \tag{1}$$

For the Continuous Sleepiness Sub-challenge, output dimension is reduced to 1, representing the level of sleepiness.

$$Y_{CS} = relu\left(zW_3 + b_3\right)W_4 + b_4, \tag{2}$$

where W_1, W_2, W_3, W_4 represent weight matrices; b_1, b_2, b_3, b_4 represent bias vectors.

2.1.2. Attention-based BLSTM

Bidirectional Long Short-Term Memory (BLSTM) has also been widely used in speech recognition tasks [17, 18]. Compared with standard LSTM, it incorporats both forward and backward information of a time series into hidden states. The input of BLSTM is frame-wise acoustic features. At each time step t, BLSTM encode the i^{th} feature f_{it} into forward and backward direction as follows:

$$\overrightarrow{h_{it}} = \overrightarrow{LSTM}(f_{it}), t \in [1, L], \tag{3}$$

$$\overleftarrow{h_{it}} = \overleftarrow{LSTM}(f_{it}), t \in [L, 1], \tag{4}$$

L is the timestep of the frame-wise feature. We can then obtain final hidden state h_{it} by concatenating $\overrightarrow{h_{it}}$ and $\overleftarrow{h_{it}}$. An attention layer is further used to reweight and encode h_{it} into context vector c. c is forwarded to a fully-connected layer and generate prediction for the Continuous Sleepiness Sub-challenge.

$$Y_{CS} = relu\left(cW_5 + b_5\right)W_6 + b_6, \tag{5}$$

where W_5, W_6, are weight matrices; b_5, b_6 are bias vectors.

2.1.3. Attention Layer

Attention is shown to be an effective technique to summarize complex information from time series. Here we use the Bahdanau attention [19] as our attention mechanism for both CNN and BLSTM mentioned above; an attentive representation is computed as a learnable weighted sum over all frames. The score function $e(\cdot)$ and attention weight α_t are defined as:

$$e\left(u_{it}\right) = v_a^T \tanh\left(W_a u_{it} + b_a\right), \tag{6}$$

$$\alpha_t = \frac{\exp(e\left(u_{it}\right))}{\sum_{t=1}^{T} \exp(e\left(u_{it}\right))}, \tag{7}$$

where $v_a \in \mathbb{R}^d$ and $W_a \in \mathbb{R}^{d \times d}$ are weight matrices, $b_a \in \mathbb{R}^{d \times 1}$ is a bias vector, d is the hidden dimension. For BLSTM+ATT, u_{it} corresponds to the outputs of BLSTM h_{it} with T equals to L. For CNN+ATT, u_{it} corresponds to the resized feature map m from a fully-connected layer after convolutional layers with T equals to C'. The attentive representation h_a is obtained by performing weighted sum over u_{it}.

$$h_c = \sum_{t=1}^{T} \alpha_t u_{it}. \tag{8}$$

2.1.4. Adversarial Autoencoder

Adversarial auto-encoder (AAE) [20] is a Generative Adversarial Networks (GAN) [21] based autoencoder approach. The AAE model includes three networks: encoder (generator), decoder and discriminator. The encoder and decoder are trained as a standard autoencoder that reconstructs features from the learned latent space. On the other hand, discriminator is trained to predict whether a sample comes from the hidden code of the autoencdoer or from a sampled distribution. In this work, we fit the sample distribution to a mixture of 10 2-D Gaussian, and each mixture represents the associated class of the Baby Sound categorization. As the training process converges, the AAE decoder can then be utilized to regenerate synthetic samples according to a class-specific code vector [20, 22]

2.2. Dataset Preprocessing

2.2.1. Styrian Dialect (SD)

Volume Augmentation: The label distribution on SD dataset is imbalanced: 4,052 for UrbanS, 1,949 for EasternS, and 1,796 for NorthernS. This imbalanced data distribution generally creates a robustness issue. In this work, we conduct data augmentation to address this problem. Data augmentation is commonly used to increase the amount of data that help avoid over-fitting. Many approaches have been explored to augment the audio data, such as adding Gaussian noise, pitch shifting, time stretch, speed change, and volume change [23]. Specifically, previous research indicates that speed change and volume change can help improve dialect recognition performances [16]. In the SD dataset, we have experimented with these two methods. However, changing speed does not improve the performance in our experiment, thus volume augmentation is used. We increase the data samples by tuning the volume of training set with the ratio of 0.5, 1, and 2, and we extract the features on this augmented set for model training.

2.2.2. Continuous Sleepiness (CS)

Select Significant Labels: Given that the property of CS dataset labels are ordinal instead of categorical classes, those data samples with moderate scores, such as 4,5,6, are highly indifferentiable and take up the majority of the dataset. In contrast, data with extreme scores, such as 1,2,3,7,8,9, are more discriminative, we assume that data with extreme scores would be more effective in learning our algorithm. We filter out those training data with moderate scores and select significant labeled data samples, named Select Significant Labels (SSL), in the CS training set for this work.

2.2.3. Baby Sound (BS)

Synthetic and Conditional Synthetic Data Augmentation

Although upsampling can reduce impact to the imbalanced class distribution issue, it does not increase the diversity, hence

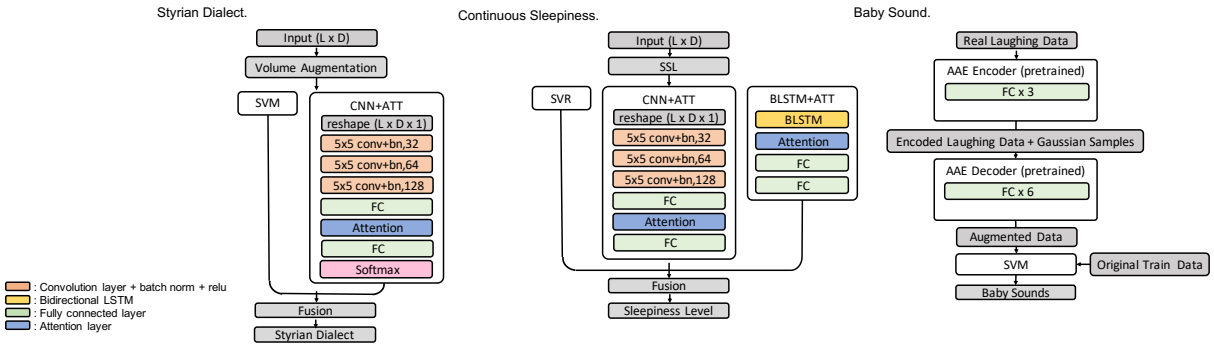

Figure 1: *Styrian Dialect: SVM and CNN+ATT are trained on augmented data are combined together in fusion stage. Continuous Sleepiness: we use SSL on training data. Predictions of SVR, CNN+ATT and BLSTM are combined in the fusion stage. Baby Sound: we utilize original train data to pretrain AAE model. Encoded laughing data vector, selected Gaussian samples and pretrain decoder are used to generate augmented samples (Synthetic feature and Conditional Synthetic feature).*

the information content, of the minority class. In order to overcome this issue, Adversarial auto-encoder (AAE) is used to regenerate synthetic samples. We sample the target code vectors from a latent mixture of Gaussian distribution of each baby sound class component. We then pass the sampled code vector through the learned decoder of the AAE to generate synthetic data. Moreover, in order to further increase the diversity of laughter data in the BS dataset, we add 30 samples of 1-second baby laughter audio data [24] as an additional data source. Instead of augmenting these raw training data directly, we input them into our learned AAE model to generate the reconstructed features (Conditional-Synthetic feature) that would be more resistant to the domain shift effect [25].

2.3. Fusion Techniques

Fusing multiple models in final prediction can not only achieve better predictive performance, but also reduce the variance and bias compared with using individual model. Despite different model architectures, we combine each model's predictive output at the decision stage using probability for classification [26] or predicted values after regression. In this work, we adopt two fusion strategies on SD and CS Sub-challenges. Additionally, instead of fusing the models' outputs at the decision-level, we conduct early fusion of feature sets on BS Sub-challenge. In SD Sub-challenge we apply decision score fusion. The probability distributions of individual model on each category are collected as an ensemble. The probability distribution of the ensemble is the average of the individual models' probability distribution. In CS Sub-challenge, we fuse the models' sleepiness regressed levels by taking the mean. Lastly, in BS Sub-challenge, different feature sets are needed to obtain an improvement in the recognition rates. In order to streamline our data augmentation process, we conduct early fusion by concatenating different feature sets to build a robust feature set.

3. Experimental Setup and Results

3.1. Experimental Setup

3.1.1. Feature Sets

Statistical Features: In this work, we extract several acoustic statistical features using the openSMILE toolkit [27], including ComParE16 functional, extended version of Geneva Minimalistic Acoustic Parameter Set (eGemaps) [28] functional and IS10 challenge [29]. Besides, baseline features, i.e., auDeep [30] and BoAW [31], are also utilized for our baseline SVM training.

Temporal Features: For time series models, we extract the low-level descriptors of eGemaps[28] (eGemaps-LLD), which is used for training spatial-temporal models such as BLSTM [18] and CNN [15]. LLD sequence is padded to the max sequence length. For CNN-based model training, we resize all eGemaps-LLD into a 3D array in the shape of $(max\ time\ steps \times dim \times 1)$, where dim is the dimension of the feature. Consequently, each LLD is treated as an image with a single channel.

3.1.2. Model Parameters

CNN+ATT: CNN+ATT combines convolutional layers with an attention layer and a dense layer in between convolution and attention, finally a fully-connected layer that learns to perform recognition on the attentive representation. The convolutional layers consist of three convolutional layers of kernel size 5×5 with batch normalization and rectified linear unit (relu) as activation. In SD Sub-challenge, CNN+ATT is optimized using Adam optimizer with cross-entropy loss as our objective. In CS Sub-challenge, CNN+ATT is optimized using mean square error as our objective.

BLSTM+ATT: Our BLSTM+ATT network consists of BLSTM with 32 hidden units per direction. It includes one attention layer and two fully-connected layers. In CS Sub-challenge, BLSTM+ATT is optimized using mean square error.

AAE: Our AAE network is composed of an encoder, a decoder and a discriminator. The encoder is composed of 3 fully-connected layers, decoder with 6 fully-connected layers, and discriminator with 3 fully-connected layers. Notice that we use relu as the activation function, and the dropout regularization probability is set with a keep-probability 0.5 for all layers except 0.75 for the last layer of the discriminator. A reparameterization step is added to the encoder. The AAE network is optimized using Adam optimizer, where the reconstruction decoder uses mean squared error and the adversarial network uses binary cross entropy as objective.

3.2. Models & Analysis

The following provides a comparison between our proposed methods and baseline models in each of the three sub-

Styrian Dialect (%)				
	SVM_{vol}	$CNN+ATT_{vol}$	Baseline	Fusion
NorthernS	50.5	68.2	43.7	57.7
UrbanS	52.1	91.1	64.3	70.0
EasternS	47.0	0.20	25.0	27.3
UAR (Dev)	49.9	53.2	44.4	**51.7**
UAR (Test)	-	-	47.0	**49.5**

Table 1: *A comparison of different models on SD Sub-challenge. SVM_{vol} and $CNN+ATT_{vol}$ represents SVM and CNN+ATT models trained with voice augmentation data.*

Continuous Sleepiness					
	SVR	BLSTM+ATT	CNN+ATT	Baseline	Fusion
ρ (Dev)	.354	.357	.354	.274	**.373**
ρ (Test)	-	-	-	.343	**.369**

Table 2: *A comparison between different models on CS Sub-challenge. All of the methods are trained with SSL.*

challenges. The performance of official baselines on development sets are presented based on the paper on the challenge [6].

3.2.1. Styrian Dialect (SD)

In SD sub-challenge, we utilize SVM and attention-based CNN models. In addition, volume augmentation is applied for all proposed models. Table 1 summarizes the performances of our experiments. By applying volume augmentation, SVM_{vol} shows a significant improvement on recognizing EasternS, i.e., recall rate of 47.0%, which is highest among all methods in the development set. Moreover, $CNN+ATT_{vol}$ that trained based on eGemaps-LLD achieves the best overall UAR of 53.2% on the development set as well as the recall rate on NorthernS (68.2%) and UrbanS (91.1%). By fusing SVM_{vol}, $CNN+ATT_{vol}$ and baseline model together, the UAR on testing set reaches 49.5%, which is 2.5% higher than the official baseline.

By examining the recall of each model, the SVM contributes the most on EasternS, and CNN+ATT contributes a high predictive power on NorthernS and UrbanS. Specifically, when comparing our model to the official baseline, recall of EasternS increases from 64.3% to 70.0% and recall of NothernS increases from 43.7% to 57.7%. Fusion result including baseline model shows a further improvement demonstrating that our proposed models are complementary to each other.

3.2.2. Continuous Sleepiness (CS)

Table 2 shows the performance of all models in CS Sub-challenge. All of the models are trained with SSL preprocessing as detailed in section 2.2.2. With the use of SSL, the prediction output can capture the extreme values more evenly; without the use of SSL, the output tends to be regressed to an average level which is 5. Both BLSTM+ATT and CNN+ATT frameworks benefit from SSL-based training, which obtains a .357 and .354 Spearman CC on the development set respectively. Moreover, using support vector regression (SVR) with SLL also achieves an improved corrleation, i.e., .354 Speaman CC.

In this work, by taking the mean from the outputs of SVR, BLSTM+ATT, CNN+ATT, and baseline, we obtain the highest Spearman CC on both development and test sets. When comparing the fused model with the baseline model, the Spearman CC of development set increases from .274 to .373. Also, Spearman CC of test set increases from .343 to .369. Through SSL training, our models help in regressing continuous sleepiness level, which is ordinal-ranking in nature and further mitigates the issue that the training distribution is heavily concentrated around the middle level.

Baby Sounds (%)				
	SVM-EF	AAE	C-AAE	Baseline
Canonical	67.2	66.7	67.7	66.4
Crying	69.9	69.9	66.8	70.6
Junk	65.2	63.1	64.4	67.3
Laughing	56.1	58.5	65.8	41.5
Non-canonical	32.5	39.6	42.5	24.1
UAR (Dev)	58.2	59.6	**61.3**	54.0
UAR (Test)	-	60.8	**62.2**	58.7

Table 3: *A comparison between models on the BS Sub-challenge. SVM-EF: early fusion of three feature set, AAE: Synthetic feature and C-AAE: Conditional Synthetic feature*

3.2.3. Baby Sound (BS)

Table 3 summarizes our BS Sub-challenge results. The baseline method achieves 54% UARs in the development set. In specifics, Early Fusion (SVM-EF), Synthetic (AAE), and Conditional Synthetic (C-AAE) data augmentation method improves the UAR to 58.2%, 59.6% and 61.3% respectively. We experimentally determine the number of augmented samples to be added to the BS dataset. We observe that the appropriate amount for data augmentation on the Canonical, Laughing, Non-canonical class that help improve the performance is 300, 200, and 800, respectively.

Generally, in an imbalanced data distribution task, we tend to augment minority classes data. But in this case, our experiments show that a better performance can be achieved if we simultaneously augment a varying number of samples on different classes. This may potentially due to the overall increase in the diversity of the database, which translates to a more robust recognition results on the test set. Furthermore, AAE model is limited in its capability to augment Laughing data due to the lack of real samples in the BS dataset. The use of additional Laughing data gathered from other data source can help increase the inherent information that AAE can model thus improve the performance. Finally, the majority classes of imbalanced data usually dominate over minority classes when using deep learning model as the classifier, in this work, we continue to use linear support vector machine that is more robust in handling this challenging classification task.

4. Conclusions

In this work, we present state-of-the-art attention-based recognition models on the three Sub-challenges of INTERSPEECH 2019 ComParE challenge. We further experiment with several data processing techniques, including volume augmentation, select significant labels and AAE-based data augmentation. Our methods outperform official baselines on the test set in each of the three Sub-challenges, specifically we obtain an UAR of 49.5%, a Spearman CC of .369 and an UAR of 62.2% on Styrian Dialect, Continuous Sleepiness and Baby Sound recognition task respectively.

5. References

[1] F. Rudzicz, R. Wang, M. Begum, and A. Mihailidis, "Speech interaction with personal assistive robots supporting aging at home for individuals with alzheimers disease," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 7, no. 2, p. 6, 2015.

[2] B. Zhou, K. Wu, P. Lv, J. Wang, G. Chen, B. Ji, and S. Liu, "A new remote health-care system based on moving robot intended for the elderly at home," *Journal of healthcare engineering*, vol. 2018, 2018.

[3] N. Mhatre, K. Motani, M. Shah, and S. Mali, "Donna interactive chat-bot acting as a personal assistant," *International Journal of Computer Applications*, vol. 140, no. 10, 2016.

[4] Q. Li, M. A. Vasarhelyi *et al.*, "Developing a cognitive assistant for the audit plan brainstorming session," 2018.

[5] S.-C. Tsai, H. Samani, Y.-W. Kao, K. Zhu, and B. Jalaian, "Design and development of interactive intelligent medical agent," in *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2018, pp. 210–215.

[6] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson10 *et al.*, "The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity."

[7] V. L. Beattie, D. R. Miller, S. E. Edmondson, Y. N. Patel, and G. A. Talvola, "Multi-dialect speech recognition method and apparatus," Feb. 2 1999, uS Patent 5,865,626.

[8] V. Beattie, S. Edmondson, D. Miller, Y. Patel, and G. Talvola, "An integrated multi-dialect speech recognition system with optional speaker adaptation," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[9] S. Peltier, "Providing culturally sensitive and linguistically appropriate services: An insider construct." *Canadian Journal of Speech-Language Pathology & Audiology*, vol. 35, no. 2, 2011.

[10] S. Javaheri and S. Redline, "Insomnia and risk of cardiovascular disease," *Chest*, vol. 152, no. 2, pp. 435–444, 2017.

[11] G. Zamzmi, R. Kasturi, D. Goldgof, R. Zhi, T. Ashmeade, and Y. Sun, "A review of automated pain assessment in infants: Features, classification tasks, and databases," *IEEE reviews in biomedical engineering*, vol. 11, pp. 77–96, 2018.

[12] H. P. Crowe and P. S. Zeskind, "Psychophysiological and perceptual responses to infant cries varying in pitch: Comparison of adults with low and high scores on the child abuse potential inventory," *Child abuse & neglect*, vol. 16, no. 1, pp. 19–29, 1992.

[13] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[14] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 801–804.

[15] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[16] S. Shon, A. Ali, and J. Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," *arXiv preprint arXiv:1803.04567*, 2018.

[17] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[18] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[20] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[22] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," *arXiv preprint arXiv:1806.02146*, 2018.

[23] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks." in *ISMIR*, 2015, pp. 121–126.

[24] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[25] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[26] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, "A survey of decision fusion and feature fusion strategies for pattern classification," *IETE Technical review*, vol. 27, no. 4, pp. 293–307, 2010.

[27] 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*. ACM, 2010, pp. 1459–1462.

[28] F. Eyben, K. R. Scherer, B. W. 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, 2016.

[29] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[30] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2017.

[31] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3370–3374, 2017.