INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Bag-of-Acoustic-Words for Mental Health Assessment: A Deep Autoencoding Approach

Wenchao Du[1], *Louis-Philippe Morency*[1], *Jeffrey Cohn*[2], *Alan W Black*[1]

[1]Language Technologies Institute, Carnegie Mellon University
[2]Department of Psychology, University of Pittsburgh

wenchaod@cs.cmu.edu, morency@cs.cmu.edu, jeffcohn@pitt.edu, awb@cs.cmu.edu

Abstract

Despite the recent success of deep learning, it is generally difficult to apply end-to-end deep neural networks to small datasets, such as those from the health domain, due to the tendency of neural networks to over-fit. In addition, how neural models reach their decisions is not well understood. In this paper, we present a two-stage approach to acoustic-based classification of behavior markers related to mental health disorders: first, a dictionary and the mapping from speech signals to the dictionary are learned jointly by a deep autoencoder, then the bag-of-words representation of speech is used for classification, using classifiers with simple decision boundaries. This deep bag-of-features approach has the advantage of offering more interpretability, while the use of deep autoencoder gains improvements in prediction by learning higher level features with long range dependencies, comparing to previous work using only low-level descriptors. In addition, we demonstrate the use of labeled emotion recognition data from other domains to supervise acoustic word encoding in order to help predict psychological traits. Experiments are conducted on audio recordings of 65 clinically recorded interviews with the self-reported level of post-traumatic stress disorder (PTSD), depression, and rapport with the interviewers.

Index Terms: acoustic word, deep learning, affective computing, social interaction, behavior markers

1. Introduction

Automated assessment of mental distress and disorders from nonverbal behaviours has shown promise in recent years [1] [2]. Prior work has make the case for speech to be a key objective for pathological traits such as depression and post-traumatic stress disorder (PTSD) [3]. The goal of this work is to find better way to predict both mental disorders of patients and their rapport with therapists in clinical sessions from audio recordings, as well as discover cues that help understand the acoustic behaviours of patients in dyadic interactions [4]. With the success of machine learning, we are particularly interested in data-driven approaches to mental assessment with an eye on the interpretability of the model used.

Bag-of-acoustic-words is an established method in acoustic modelling and classification tasks, such as event detection [5, 6, 7] and emotion recognition [8]. It relies on clustering to construct a codebook of acoustic words, and use the quantization of audios for classification. It earned its popularity for it extracts features in an unsupervised manner without human knowledge, while still performs reasonably well. Representing utterances and interview sessions as bag-of-words has several advantages for our specific problem. First, the extraction of acoustic words may capture localized features in human speech that are indicative of pathological behaviours. Second, combined with linear classifiers, bag-of-word representation offers more interpretability, since the importance of each word is measured by its weight. Consequently, it helps pinpoint the most informative moments in speech.

Prior work [9] has shown deep neural networks may learn useful representation of acoustic units or segments. Recently, there has been interests among deep learning community in combining bag-of-features methods with neural models. Passalis et al.[10] proposed a bag-of-features layer to perform k-means clustering of feature maps with gradient descent, and use the Gaussian encodings of features for final classification. Brendel et al.[11] proposed an architecture in which image classification is performed on small local patches, and the heat maps of local patches are aggregated for final prediction.

We believe using deep autoencoder for learning representations of acoustic words has the following advantages. First, incorporated with recurrent neural networks (RNN), neural models can potentially capture long range dependencies and embed contextual information into acoustic words. Second, the construction of codebook and the quantization of speech signals are unified in an end-to-end framework. Third, it is easy to leverage labeled data to supervise the encoding of acoustic words. The novelty of our work is two-fold: to our knowledge, we are the first to combine the bag-of-words method with deep autoencoders; and a transfer learning technique through acoustic units to make the task of mental health assessment benefit from related emotion recognition tasks.

2. Dataset

Our experiments are performed on the Distress Assessment Interview Corpus [12]. There are 65 human-to-human sessions in total. PTSD and depression are assessed with PTSD Checklist Civilian version (PCL-C) and the Patient Health Questionnaire, depression module (PHQ-9) [13]. PCL and PHQ are discretized, and the thresholds are set to be 34 and 9, respectively. Scores above the thresholds are considered positive. The correlation between PTSD and depression is 0.62. Rapport is estimated from a questionnaire as described by Gratch et al. [14]. Answers are weighted according to the nature of the questions and summed up. The median of the scores is set as threshold, and participants with scores above the threshold is treated as positive, otherwise negative. Among the 65 participants, 27 have PTSD and 21 have depression.

We use IEMOCAP [15] dataset as the source to transfer emotion recognition knowledge. 10 actors were recorded in 5 sessions (2 actors each). They were asked to perform hypothetical scenarios, both scripted and improvised, with elicit emotion contents. The emotions were annotated by 4-6 native speakers with majority vote. We adjust the labels by randomly choosing

http://dx.doi.org/10.21437/Interspeech.2019-3059

from one annotator if there is no label prevails (cannot decide), or if the label is "other". Again, only audio recordings are used for training.

3. Methodology

3.1. Feature Extraction

We consider three sets of features. Spectral: 39 dimensional Mel-Fourier Cepstral Coefficients (MFCC) consisting of 12 MFCC and raw energy, their deltas and delta-deltas. Prosodic: 3 dimensional features including fundamental frequency, voicing probability and loudness. These two sets are extracted using openSMILE [16] with 25 ms long frames and 10 ms frame rate. Voice quality: 75 dimensional features extracted using COVAREP [17], including spectral envelope, sinusoidal, glottal flow, and phase-based features.

3.2. Architecture

The architecture of the encoder is similar to the common ones used in speech recognition [18] at a high level. A convolutional neural net is first introduced to downsample and transform the signals, $\mathbf{X} \in \mathbb{R}^{c \times l}$. It outputs a sequence of feature maps, each of which corresponds to a fixed-sized segment in the input audio:

$$f_{cnn} = CNN_{enc}(\mathbf{X})$$

Then a recurrent layer is used to incorporate contextual information into each feature map. A linear map is used to reduce the dimension of input feature maps before feeding into the RNN:

$$f_{rnn} = RNN_{enc}(\mathbf{W_e} f_{cnn})$$

Then the hidden states of the RNN is projected onto the probability simplex with softmax function to obtain the distribution of acoustic words:

$$\mathbf{W} = \sigma(\mathbf{P} f_{rnn})$$

The decoder is mostly the "reverse" of the encoder. Acoustic words are embedded and fed into another RNN:

$$g_{rnn} = RNN_{dec}(\mathbf{EW})$$

The hidden states of the decoder RNN is then linearly projected and fed into a deconvolutional neural network:

$$\mathbf{Y} = CNN_{dec}(\mathbf{W_d} g_{rnn})$$

The final output, $\mathbf{Y}$, is the reconstruction of the input signal. The autoencoder is trained by minimizing the squared l_2 reconstruction loss, i.e. $||\mathbf{X} - \mathbf{Y}||_2^2$.

We now elaborate on the details of the convolutional and deconvolutional networks. We use 1d convolution over time domain in both encoder and decoder. We use strided convolution in encoder and transposed strided convolution in decoder, as in [19]. We do not use any pooling operation, as strided convolution can achieve similar downsampling effect, with less memory usage. Rectified Linear Unit is applied after each convolutional layer, before batch normalization [20]. More details of the architecture are listed in Table 1.

In our approach, segmentation of audio signals is realized through the convolutional neural network. At layer i, the length l_i and step size t_i of the corresponding audio segment of each feature map are decided by the sizes of filters k and stride s of previous convolutional layers, and are given by the following recurrences:

$$l_i = l_{i-1} + (k_i - 1) * t_{i-1} \quad (1)$$

$$t_i = t_{i-1} * s_i \quad (2)$$

with starting conditions $l_0 = 1$ and $t_0 = 1$.

Table 1: *Architecture of the autoencoder. K is the vocabulary size. D is the dimension of data. At each layer, the first parameter is the input dimension and the second the output dimension. Dropout [21] is applied to the inputs of the convolutional layers and both the inputs and outputs of the recurrent layers.*

Encoder	Decoder
Linear(256, K)	Linear(K, 128)
GRU(128, 256)	GRU(128, 128)
Linear(256, 128)	Linear(128, 256)
BatchNorm()	
Conv(128, 256, k=5, s=2)	TransConv(256, 128, k=5, s=2)
BatchNorm()	BatchNorm()
Conv(64, 128, k=5, s=2)	TransConv(128, 64, k=5, s=2)
BatchNorm()	BatchNorm()
Conv(32, 64, k=5, s=2)	TransConv(64, 32, k=5, s=2)
BatchNorm()	BatchNorm()
Conv(D, 32, k=5, s=2)	TransConv(32, D, k=5, s=2)

3.3. Transfer Learning

To leverage labeled data, we add a classifier to the top of the encoder. More specifically, for each audio signal, we sum up the distribution of acoustic words of all its segments (i.e. sum up $\mathbf{W}$ along time axis), and use the resulting bag-of-words representation, $\mathbf{w}$, to classify its emotion label, L.

$$P(L(\mathbf{X})) = \sigma(\mathbf{L_c w})$$

The rest part of the model shares parameters with that trained on the interview sessions. The model is trained with the sum of the reconstruction loss and weighted cross entropy loss with respect to labels. To avoid being dominated by the external domain, data from the labeled domain are sampled with smaller batch size than the local domain during mini-batching.

3.4. Training Details

All models are implemented in PyTorch [22]. Half of the data from each speaker are held out for validation during training the autoencoder. We use Adam optimization algorithm [23] with learning rate 0.0001 and exponential decay rate of 0.9. We use dropout rate of 0.5. The training stops after 30 epochs. We use batch size of 16. For transfer learning, IEMOCAP data are sampled with batch size of 8. The reconstruction and classification loss of IEMOCAP data are down-weighted with 0.1.

4. Experiments

4.1. Main Results

We now describe the baselines. The first baseline is simply averaging all the frames in each session. The other baselines use k-means clustering to construct the codebook. In addition to frame-level (LLD) acoustic words, we also consider windowed acoustic words. 10 consecutive frames of features are concatenated as a window and clustered, and we set the step size of

Table 2: *Main results at vocabulary size of 25.*

		PTSD		Depression		Rapport	
		Accuracy	F1	Accuracy	F1	Accuracy	F1
MFCC	**Averaging Frames**	0.6250	0.5200	0.6094	0.4186	0.5000	0.5000
	LLD Acoustic Words	0.5000	0.5294	0.5781	0.5263	0.3667	0.4722
	Windowed LLD Acoustic Words	0.4844	0.5823	0.5156	0.5231	0.3667	0.4412
	Neural Acoustic Words	0.6563	0.6207	**0.6719**	**0.5333**	0.5000	0.5714
	Supervised Neural Acoustic Words	**0.6563**	**0.6452**	**0.6563**	**0.5600**	**0.5238**	**0.5714**
Prosodic	**Averaging Frames**	0.5156	0.2791	**0.6094**	**0.5098**	0.5667	0.1333
	LLD Acoustic Words	0.5156	0.4918	0.4844	0.5479	0.3666	0.5000
	Windowed LLD Acoustic Words	**0.5781**	**0.5263**	0.4688	0.5405	0.4167	0.5205
	Neural Acoustic Words	0.5781	0.5091	0.5625	0.5000	**0.4333**	**0.5278**
	Supervised Neural Acoustic Words	0.5625	0.5172	**0.5781**	**0.5263**	0.4167	0.5070
Voice Quality	**Averaging Frames**	0.4531	0.4068	0.5781	0.4906	0.6167	0.4103
	LLD Acoustic Words	0.5938	0.5806	0.6094	0.5283	0.5167	0.4727
	Windowed LLD Acoustic Words	0.5938	0.5667	**0.6406**	**0.5818**	0.5000	0.4444
	Neural Acoustic Words	0.6094	0.5614	0.6250	0.5200	0.4167	0.5205
	Supervised Neural Acoustic Words	**0.6563**	**0.6071**	0.6406	0.5106	**0.4667**	**0.5428**

windows to 5. As for quantization method, hard assignment has been previously shown to often yield inferior results [24]. Since our deep autoencoding approach inherently learns soft assignment of words, for the sake of fair comparison, we experimented with two alternative assignment methods for the baselines: (i) soft assignment, in which scaled squared Euclidean distances to centroids $\mathbf{m_i}$ are normalized using softmax:

$$\mathbf{h_i} = \frac{\exp(-||\mathbf{x} - \mathbf{m_i}||^2/\tau)}{\sum_i \exp(-||\mathbf{x} - \mathbf{m_i}||^2/\tau)}$$

and (ii) multiple assignment, in which the top few closest words are assigned. In our experiments, we found that soft assignment generally works better, so we use it in all our baselines. τ is set to be the square root of vocabulary size.

For all the methods based on bag-of-words, we tried two types of featurization and report the best results: one is using raw count of acoustic words, and one is using log smoothed count: $\log(1 + x)$. We use Gaussian Naive Bayes for classification, and we found it generally performs better than support vector machine on our dataset. Prediction performances are evaluated using held-one-speaker-out training and testing. All input features are z-normalized. When extracting acoustic words (both clustering and autoencoding), audios from both participants and interviewers are used. Audios from participants only are used for classification. The initialization of k-means clustering is done through k-means++ [25]. The full algorithm is implemented using scikit-learn [26].

The main results are shown in Table 2. For each set of features and task, the best results are highlighted. We consider F1 score the most important and accuracy as secondary. The overall best results are given by our neural acoustic words from MFCC features, supervised or unsupervised. Generally speaking, supervision can improve the performance regardless of the features used, except in one case where it hurts predicting rapport using prosodic words. In some cases, LLD acoustic words can perform better than or comparably well as neural acoustic words, but we notice that in such cases the LLD words usually are much worse on other tasks. In other words, neural acoustic words are more well-rounded than LLD acoustic words. The other thing worth noting is that for bag-of-words approaches, MFCC are generally better than voice quality features, and voice quality in turn better than prosodic features.

To study the effect of length and step size of acoustic words without changing the architecture, we drop every n^{th} frame in the data. As a results, the length and step size of words increase by $\frac{1}{n-1}$. We experimented with $n = 2$ and 3 for each set of features. As shown in Table 3, longer words and steps do not help prediction in general.

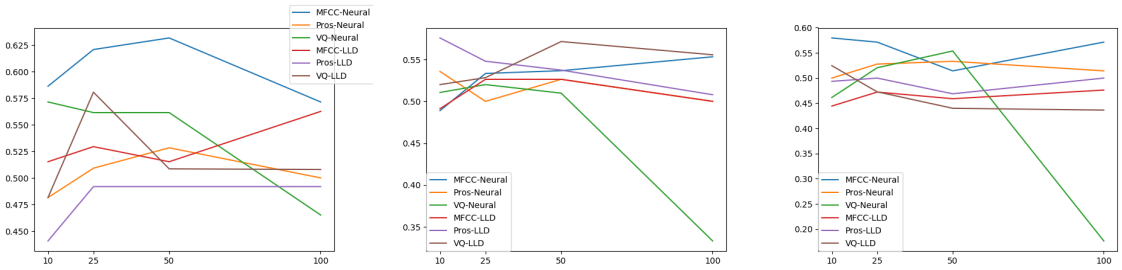
We also studied the effect of vocabulary size for clustering and autoencoding methods, and the results are plotted in Figure 1. Acoustic words from voice quality features see significant drop in F1 scores when the vocabulary size reaches 100. In the case of depression, prosodic words (both clustering and autoencoding) perform the best when the vocabulary size is the smallest (10), and degrade when the size becomes larger.

Table 3: *F1 scores when varying length and step size at vocabulary size of 25 for each set of features.*

	PTSD	Depression	Rapport
MFCC	0.6207	0.5333	0.5714
MFCC (half longer)	0.5902	0.5490	0.5634
MFCC (double)	0.5455	0.5405	0.5321
Prosodic	0.5091	0.5	0.5278
Prosodic (half longer)	0.5091	0.4912	0.5505
Prosodic (double)	0.5091	0.5	0.5143
VQ	0.5614	0.52	0.5205
VQ (half longer)	0.5667	0.5385	0.4262
VQ (double)	0.5455	0.4898	0.5507

4.2. Ablation Study

To study how useful contextual information is, we replace the recurrent layers with a linear layer with tanh activation, so acoustic words are learned without contexts. As can be seen in Table 4, performance is slightly worse than using RNN for MFCC, but marginally better for prosodic features. As for voice qualities, removing RNN hurts the prediction of PTSD and de-



(a) *PTSD* (b) *Depression* (c) *Rapport*

Figure 1: *F1 scores when varying vocabulary size for bag-of-acoustic-words methods.*

pression while helping rapport. In summary, the usefulness of modelling context is not only feature-specific but also task-specific.

Table 4: *Model ablation studies with RNN taken out. Vocabulary size is 25. F1 scores are listed in the table.*

	PTSD	**Depression**	**Rapport**
MFCC	0.6207	0.5333	0.5714
- RNN	0.5574	0.5106	0.5333
Prosodic	0.5091	0.5	0.5278
- RNN	0.5283	0.5263	0.5333
VQ	0.5614	0.52	0.5205
- RNN	0.5490	0.3256	0.6000

4.3. Qualitative Analysis

To understand what the deep autoencoder is discovering about the ways of people speaking, we investigate what linguistic words and phrases are associated with those acoustic words predictive of mental disorders and rapport. We use Montreal Forced Aligner [27] with transcriptions to obtain the word boundaries. We align the boundaries of acoustic words to the closest boundary of linguistic words in utterances, so that acoustic words are aligned with the n-grams in transcriptions. We perform one tailed t-test on bag-of-acoustic-words representations of sessions to decide the most predictive acoustic word for depression, PTSD, and rapport separately. We then list the linguistic n-grams that have most of these acoustic words on average and are spoken by at least two speakers.

The top 20 words for each psychological trait is listed in Table 5. The most representative phrases are highlighted. The phrases under PTSD has more first pronoun and fillers; negative tones are common. The phrases under depression share a large subset with PTSD and have a similar nature. The expressions under rapport tend to be more about specific things and activities, and less self-referential.

5. Conclusion

We studied automated assessment of mental disorders and rapport of humans from acoustics. We developed a novel deep learning approach for feature extraction and representation, combined with bag-of-words approach for classification. The deep autoencoding approach is shown to be competitive to previous clustering-based approaches. We also showed the possibility of leveraging emotion labels from other dyadic interaction data to help predicting mental distress and rapport. In our quantitative analysis, we found that MFCC features work the best together with bag-of-words method, while prosodic and voice quality features are proven to be useful as well. One future direction would be finding a better fusion method for combining features. In our preliminary qualitative analysis, we showed that the acoustic words learned from the deep autoencoding approach exhibited interesting patterns and correlation with linguistic cues for each of the mental assessment traits, thus providing insights on the link between human acoustic behaviours and psychology. Our quantitative evaluation also shows that automated mental state assessment from acoustics is a challenging task and there is still plenty of room for improvement.

Table 5: *Top linguistic n-grams that most often coincide with the most predictive acoustic words of different sets of features.*

PTSD	**i've never**, south, happened i, and they have, should, i actually, **just uh**, alive, **and i didn't**, **uh i have**, i made a, stuff to, hmm, it was in, **i don't know i**, **but it was**, **no i have**, yeah she, i did a, where would i
Depression	**i've never**, happened i, and they have, south, yeah she, and i think that's, i actually, **and i didn't**, **i hate**, i made a, stuff to, it was in, i learned, **no i have**, a man, **uh huh**, yeah so, well i'm, i mean she, oh that
Rapport	there's two, it i mean, i just wanna, they'd, look for, **my sister**, **to school**, and stuff and, seeing a, **his face**, says she, to him, **drunk**, got some, **michigan**, and got, the help, i was still

6. Acknowledgements

This material is based upon work partially supported by the National Science Foundation (IIS-1721667 and IIS-1722822). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation, and no official endorsement should be inferred.

7. References

[1] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, "Multimodal assessment of depression from behavioral signals," in *The Handbook of Multimodal-Multisensor Interfaces*. Association for Computing Machinery and Morgan & Claypool, 2018, pp. 375–417.

[2] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD," in *INTERSPEECH*, 2013, pp. 847–851.

[3] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[4] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, 2013.

[5] S. Rawat, P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze, "Robust audio-codebooks for large-scale event detection in consumer videos." in *INTERSPEECH*, 2013, pp. 2929–2933.

[6] A. Plinge, R. Grzeszick, and G. A. Fink, "A bag-of-features approach to acoustic event detection," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3704–3708.

[7] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using lbp-hog based bag-of-audio-words feature representation," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech." in *Interspeech*, 2016, pp. 495–499.

[9] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *Interspeech 2016*, pp. 765–769, 2016.

[10] N. Passalis and A. Tefas, "Learning bag-of-features pooling for deep convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5755–5763.

[11] W. Brendel and M. Bethge, "Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet," in *International Conference on Learning Representations*, 2019.

[12] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, and L.-P. Morency, "The Distress Analysis Interview Corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, May 2014, pp. 3123–3128.

[13] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[14] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, "Creating rapport with virtual agents," in *International workshop on intelligent virtual agents*. Springer, 2007, pp. 125–138.

[15] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[16] 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*. ACM, 2013, pp. 835–838.

[17] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 960–964.

[18] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4580–4584.

[19] Y. Zhang, D. Shen, G. Wang, Z. Gan, R. Henao, and L. Carin, "Deconvolutional paragraph representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4169–4179.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] S. Pancoast and M. Akbacak, "Softening quantization in bag-of-audio-words," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1370–1374.

[25] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[27] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, 2017.