INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Empirical Interpretation of Speech Emotion Perception with Attention Based Model for Speech Emotion Recognition

Md Asif Jalal, Rosanna Milner, Thomas Hain

Speech and Hearing Group (SPandH), The University of Sheffield
m.a.jalal, rosanna.milner, t.hain @sheffield.ac.uk

Abstract

Speech emotion recognition is essential for obtaining emotional intelligence which affects the understanding of context and meaning of speech. Harmonically structured vowel and consonant sounds add indexical and linguistic cues in spoken information. Previous research argued whether vowel sound cues were more important in carrying the emotional context from a psychological and linguistic point of view. Other research also claimed that emotion information could exist in small overlapping acoustic cues. However, these claims are not corroborated in computational speech emotion recognition systems. In this research, a convolution-based model and a long-short-term memory-based model, both using attention, are applied to investigate these theories of speech emotion on computational models. The role of acoustic context and word importance is demonstrated for the task of speech emotion recognition. The IEMOCAP corpus is evaluated by the proposed models, and 80.1% unweighted accuracy is achieved on pure acoustic data which is higher than current state-of-the-art models on this task. The phones and words are mapped to the attention vectors and it is seen that the vowel sounds are more important for defining emotion acoustic cues than the consonants, and the model can assign word importance based on acoustic context.

Index Terms: speech emotion recognition, speech emotion intelligibility, computational paralinguistics

1. Introduction

The aim of speech emotion recognition (SER) is to automatically detect human emotions from spoken audio [1, 2] and research in vocal expression recognition of emotion is interdisciplinary. There have been several reviews in the field [3, 4, 5, 6] and previous research in psychology has attempted to represent emotions using different models, such as Plutchik's wheel of emotion [7] or the hourglass of emotions [8]. The ground truth is hugely dependant on the listeners who associate the acoustic cue patterns with discrete emotion states. However, emotions are complex as they cannot be clearly defined, which makes it difficult to detect them accurately. Two major questions arise about these acoustic cues. The first is concerning acoustic elements in the cues and the second investigates the length and boundaries of the cues.

The most fundamental distinction that can be made in speech sounds is between vowels and consonants [9, 10]. These sounds also carry socio-linguistic information, and previous research on phonetics and psychology argue whether vowel or consonant sounds are more dominant in determining the underlying emotion. Previous research suggests that consonants play a more vital role in delivering socio-linguistics information. However, most of the recent research shows that vowels play the vital role [11, 12]. Vowels have higher variation in formant structures, allowing them to be more enriching in acoustic context [9]. These claims have not been corroborated in computational speech emotion recognition systems, and it is not typical for current deep neural models to be interpreted with speech emotion data in this way.

On the contrary, from a computational point of view SER tasks require a front-end for extracting features that hold maximum correlation with emotion attributes while being robust to changes in time, frequency, speaker, medium and other external distortions. SER systems train a classifier or a group of classifiers to map speech data to a categorical distribution of different emotions. In practice, the most popular features are Opensmile [13], eGeMaps [14], MFCCs [15] and filterbanks [16]. These features are used with different classifiers such as hidden Markov models (HMMs) [17], support vector machines (SVMs) [18], deep belief networks (DBNs) [19] and deep neural networks (DNNs). DNNs learn task-specific abstract feature representations by filtering out unnecessary information and improving generalisation [20, 21, 22]. Research has proposed representation learning by modelling mid to long-term sequence dependencies [23, 24, 25].

Most recently, [26] presented a domain adversarial system for investigating whether information in acted datasets can be learnt to benefit emotion prediction for natural datasets. The work aimed to be consistent by only considering datasets with adult English speakers with the big-six emotions: happiness, sadness, anger, surprise, disgust and fear. The method applies a bi-directional long short-term memory (BLSTM) with an attention layer and trains in a domain adversarial fashion. It uses sequence modelling, which is arguably more appropriate for use with emotions that change over time. Alternatively, [27] presented a convolution-based self-attention (CSA) model for speech emotion recognition with fixed-length context sizes. Both the models achieved very high accuracy compared to the current state-of-the-art models.

In this work, two computational SER models based on our previous work [26, 27] have been interpreted in light of the proposed phonetic, linguistic and psychological claims about the acoustics cues for speech emotion recognition in humans. It is shown that the attention weights in the proposed networks are hugely inclined to the vowel sounds, and it imposes word importance based on preceding/following acoustic context and prosody. It is also shown that smaller acoustic contexts are vital in carrying emotions as previously hypothesised.

2. Consonant-Vowel Boundaries and Perception

Traditionally, it was observed that consonant sounds carry the more important speech information until recent studies questioned this claim [11, 12, 28]. Cole et al. [11] and Fogerty et al. [12] have found that among humans, vowel-only segments have higher intelligibility at sentence level stimuli than

http://dx.doi.org/10.21437/Interspeech.2020-3007

consonant-only segments. Owren et al. [29] performed a similar experiment on word-level stimuli which mostly agrees with [11], that meaning is more comprehensible to listeners with vowels. Furthermore, [29] adds that vowel phones at the beginning of a word constitute to improved understanding for listeners, but consonant cues add more acoustic context for meaning. In sentences, vowels occur 10% less than consonants [30] but the vowel proportions yield the maximum intelligibility in a sentence.

It is a fact that the acoustic cues for sentence intelligibility are distributed across consonant-vowel boundaries [12]. The perceptual cues associated with the acoustics in terms of vowels and consonants interact with each other to build an acoustic-phonetic context for speech perception [31, 32]. So, these small overlapping cues also hold some portion of perceptual and socio-linguistic information though it is not clear whether these types of acoustic cues can be useful for computational SER tasks.

The role of vowels in perceiving socio-linguistic information and emotion can also be explained with the different harmonic structures of vowels [9]. These different harmonic variations in long periods of time constitute prosody which plays a vital role in delivering emotion. Warama et al. [33] used short vowel samples (150 ms) to remove the prosody effect of vowels and showed that it is possible to perceive emotion from shorter utterances. This implies SER systems could be trained on shorter segments or parts of longer utterances.

3. DNN Approaches

To investigate the attention on the phones and the importance of short segments in regards to SER, two audio-only state-of-the-art approaches proposed in SER are considered. The first is based on sequence modelling and is trained using a bidirectional long short-term memory network (BLSTM) [26], and the second is trained using convolutional neural networks (CNNs) [27] and is not a sequence model. The BLSTM model has been used to investigate the attention put on phones with sentence-level sample. The CNN based model is used to investigate the importance of the short overlapping acoustic cues for SER.

3.1. BLSTM with attention (BLSTMATT)

This approach applies a BLSTM followed by an attention layer and has been described in detail in [26]. LSTM networks ignore the future context and rely on the temporal order of the sequence, whereas BLSTMs [34] introduce a second layer of hidden connections which flows in the opposite temporal direction as a way to exploit the contextual information from the past and the future [35]. Applying these networks, a temporal feature distribution over the sequence can be obtained, which is useful for SER tasks.

Attention has the flexibility of computing long-term inter-sequence dependencies. By computing the global mean, the attention mechanism focuses the network onto specific parts of itself which in turn captures global information. The non-linearity $tanh$ is used to multiply the global mean over the whole temporal vector which computes the positional dependency of each element. The resulting vector is used to compute the attention weights using $softmax$. The soft attention mechanism is also adopted for this work and the multiplicative method is applied as in [36].

Finally, the classifier stage of the network contains a fully connected linear layer which projects the attention output down to the number of emotions present. It passes through a $softmax$ layer before computing the loss.

3.2. Convolutional Self-Attention (CSA)

The approach in [37] extracts a spatial feature y using a CNN and performs task-specific high dimensional feature expansion using a self-attention network, which is projected to the original feature dimension. The new feature $\hat{y}$ will be

$$\hat{y} = y + \gamma(A) \tag{1}$$

where the learnable parameter γ controls the degree of projection and A is the attention map.

Convolution layers with smaller kernel sizes (2-6) have been applied. The features from previous CNN layers, $\mathbf{y}$, are transformed into seven feature spaces in Eq. 2 and Eq. 4.

$$j\left(y\right) = \boldsymbol{W_j}\left(y\right) \qquad\qquad k\left(y\right) = \boldsymbol{W_k}\left(y\right) \tag{2}$$

where j and k are feature spaces learned through the convolutional layers $\boldsymbol{W_j}$ and $\boldsymbol{W_k}$. The positional relationship between the elements in j and k are calculated. This is followed by calculating the attention energy, E.

$$E = \mathit{softmax}\left(j\left(y\right)^T k\left(y\right)\right) \tag{3}$$

The attention energy is projected onto a common representation space l.

$$l\left(y\right) = \boldsymbol{W_l}\left(y\right) \tag{4}$$

where $\boldsymbol{W_l}$ is a convolutional layer and the network weights are learned through back-propagation. Each of these projections, except l, performs downsampling of the input feature maps.

The attention map, A, is calculated by performing matrix multiplication as shown in Eq. 5 and the attention is projected into the same dimension as the original feature, y. The projection is controlled using γ.

$$A = \gamma \cdot \left(l\left(y\right) \cdot E\right) \tag{5}$$

This network learns the non-local dependencies as well as the local neighbourhood using the convolution self-attention.

4. Experimental Setup

4.1. Data

The IEMOCAP [38] dataset has been used for evaluation. The corpus comprises over 12 hours of utterances from 10 speakers (5 male and 5 female) [38]. It has five dyadic (between two speakers) sessions, and the sessions are either scripted or improvised for eliciting emotions. The spoken English has a North American accent and in previous research it is common for IEMOCAP to be evaluated as four classes only: *happy* (*happy* is combined with *excitement* to give 1545 segments), *sad* (1084 segments), *anger* (1103 segments) and *neutral* (1708 segments). The utterances are split into a train set of 4290 (Sessions 1-4) and a test set of 1241 (Session 5) and referred to as IEM4 in this paper and in [26].

4.2. Features

Experiments in [26] showed how the *BLSTMATT* system performed best in terms of unweighted and weighted accuracy with 23-dimensional log-Mel filterbank features which are applied to the *CSA* system as well.

4.3. Implementation

The two systems are implemented in PyTorch [39]. The *BLSTMATT* performs segment-level classification and the *CSA* performs frame-level classification. The Adam optimiser [40] is applied to the two models with the initial learning rate of 0.0001. As Adam adaptively optimises the learning rate but does not change it, the PyTorch approach of ReduceLROnPlateau has been investigated. The optimum patience setting is found to be 4 epochs with a multiplicative factor of 0.8. System combination is also explored to investigate whether these models have a complementary relationship for SER.

4.3.1. Segment Level

The *BLSTMATT* contains two hidden layers of 512 nodes each. The output layer of size 1024 is fed into the attention mechanism computing a context vector of size 128, which is projected to 1024 nodes. This is then passed to the emotion classifier which linearly projects to the 4 classes. The cross-entropy loss function is applied, which is preceded by a $softmax$ layer in the PyTorch implementation. The *BLSTMATT* produces a variable length attention vector based on the input segment length, as mentioned in section 3.1. To interpret the acoustic attention, the attention vectors have been extracted and mapped with the phones in the input segments.

4.3.2. Frame Level

The *CSA* consists of three CNN blocks, each block has batch normalisation and rectified linear unit (ReLU) activation. These layers produce 128 channel feature maps which are fused in a convolutional self-attention layer where the number of channels are downsampled. The contextually enhanced output features from the attention layer are given as input to the classifier which linearly projects to the 4 classes.

To investigate the acoustic context length, the utterances are split into chunks with an overlap of 10 frames. The utterances which are less than the context length are not included in the training or test sets. The size of the chunks is varied from 20 to 120 frames.

4.4. Evaluation

Unweighted accuracy (UA) and the weighted accuracy (WA) are used to evaluate the results. The UA calculates accuracy in terms of the total correct predictions divided by total samples, which gives equal weight to each class. As IEM4 is imbalanced across the emotion classes, the WA is calculated as well, which weighs each class according to the number of samples in that class:

$$UA = \frac{TP + TN}{P + N}, \qquad WA = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N}) \qquad (6)$$

where P is the number of correct positive instances (equivalent to $TP + FN$) and N is the number of correct negative instances (equivalent to $TN + FP$).

4.5. Baseline

The results are directly compared with other SER systems which also use the IEM4 dataset and process only audio. For WA, [41] applies factor analysis in a cross-lingual approach. For UA, in [42] a CNN-LSTM model is trained, [24] applies a deep capsule network with gated recurrent units (GRU) for sequence modelling, [43] used deep attention pooling for SER tasks and [43] applies attention pooling. Finally, a multimodal system which is also attention-based and processes both audio and textual data [44] is included to show the performance the presented audio-only systems could achieve.

System	Context	UA%	WA%
Factor analysis [41]	-	-	56.1
CNN_LSTM [42]	-	59.4	-
CNN_RecCap [24]	-	58.1	-
CNN_GRU-SeqCap [24]	-	59.7	-
Attention Pool [43]	-	71.8	-
MULTIMODAL: Attention [44]	-	*78.0*	-
BLSTMATT	Variable	**80.1**	**73.5**
CSA	20	75.8	**69.4**
	30	**76.3**	68.8
	40	75.1	68.0
	50	73.9	67.8
	60	75.1	67.0
	70	74.1	64.7
	80	73.2	67.4
	90	74.8	66.9
	100	74.6	65.9
	110	73.8	67.5
	120	72.2	64.2
SYSCOMB: *BLSTMATT* with *CSA*	V./30	**80.5**	**74.0**

Table 1: *Results for both model architectures and system combination compared to baseline results on IEM4 data.*

5. Results and Discussion

The experimental results are shown in Table 1. The *BLSTMATT* system is trained and tested with whole segments from the corpus. Naturally, the context length for *BLSTMATT* is variable because the segment lengths are not fixed in IEMOCAP. On the contrary, the *CSA* system is trained with fixed-length samples.

The *BLSTMATT* system outperforms the baselines in terms of UA and WA on IEM4. It even outperforms the multimodal system which makes use of textual information as well as audio, which the two presented models do not use. The *BLSTMATT* system outperforms the *CSA* model by 2.7% absolute difference. One of the possible reasons is that the *BLSTMATT* is trained with the whole segment, taking in all the information possible. Typically, in emotion recognition corpora a whole segment is labelled as one emotion category. However, all the smaller acoustic cues from the segment don't necessarily belong to the same emotion category because emotions are dynamic entities and can change momentarily. This segment issue has been discussed later.

The *CSA* system outperforms the best baselines in terms of UA and WA. However, unlike the *BLSTMATT*, it does not outperform the multimodal [44] baseline. When comparing the context lengths, *CSA* shows better performance with smaller context lengths, and the best result of UA 76.3% comes with context length of 30. This result does not mean that acoustic length 30 is the optimal acoustic cue length because this particular result is based on the model architecture. However, it can be clearly said that the smaller acoustic cues hold socio-linguistic emotion information as previously claimed by the cognitive studies.

The two best system outputs (*BLSTMATT* and *CSA* with context length 30) can be combined to investigate whether a gain can be achieved from the different training methods. With the *CSA* output posterior probabilities scaled by a factor of 0.4 and then multiplied by the posterior probabilities from the *BLSTMATT* gives a gain of 0.4% UA and 0.5% WA. This shows both systems learn the emotion classes in different ways, leading to overall improved performance when combining system

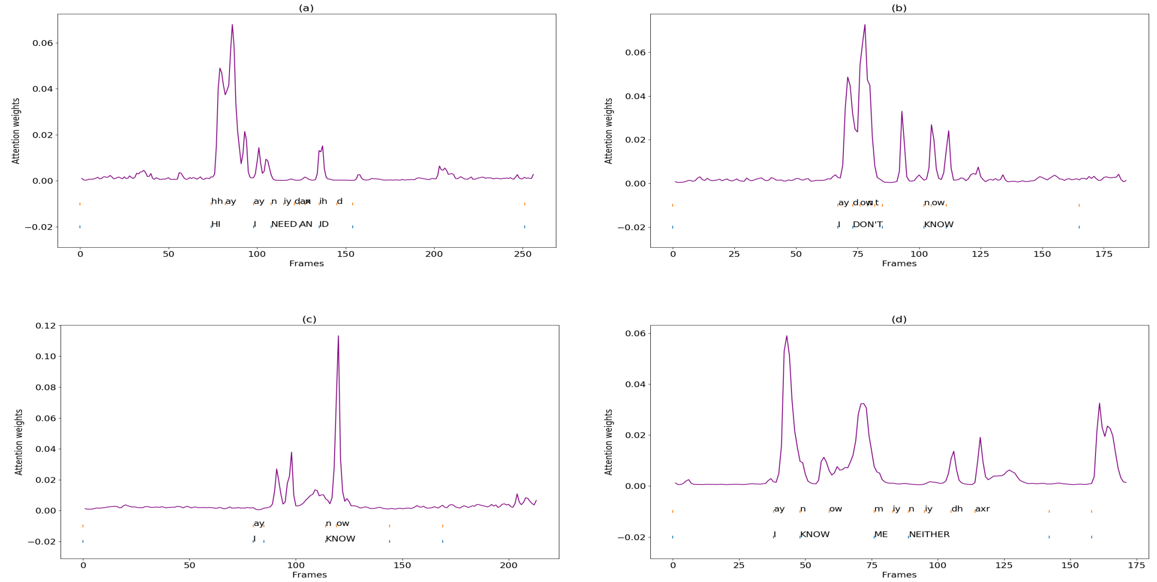

Figure 1: *Acoustic attention weights on four different segments from left to right (a) Neutral: "Hi, I need an ID", (b) Sad: "I don't know", (c) Sad: "I know" and (d) Happy: "I know, me neither"*

outputs.

Further remarks in regard to emotion classifcation for the SER task, the trained neural network classifiers map the input sample to a categorical distribution. Therefore, the output of the supervised DNN SER classifiers is based on the ground truth provided by the annotators. It is likely that some emotions are redundant and challenging to infer based on the different voiced emotion portrayals across cultures. Also, perceptual differences can clearly be seen among the manual annotators of ground truth in IEMOCAP. Speech emotion is a continuous and dynamic process, and it is not logical to consider an emotion state over a long segment. So, using small overlapping acoustic cues for determining emotion state would be a pragmatic future path.

5.1. Attention to acoustic cues

For the *BLSTMATT* system, the attention weights for each test segment can be extracted and plotted against the aligned segment. These plots are shown in Figure 1. The phones are mapped to the attention vectors to show the relative positions of the attention weights compared to the phones and words. The segments displayed have some common words, but each segment falls under a different emotion category. Common words have been investigated from different emotion categories to demonstrate the word importance weights given by the attention vectors.

Firstly, it can be seen that the attention weights are higher and prominent near the vowel phones, which implies the vowels are incredibly significant for speech emotions. There is a strong correlation seen between vowels and high attention weights. The attention weights on the consonant phones are not high, but they are not negligible. The attention vector projections on consonants are dependent on the vowel cues, and they constitute consonant-vowel boundaries in the context of emotion. These figures show similarity with the hypotheses and the claims about vowels and emotions from phonetics, psychology and linguistic studies mentioned in Section 2.

Secondly, the model gives an idea about the word importance in determining an emotion class. Here *word* is used from an acoustic point of view as the *BLSTMATT* model has not been trained with any language model. For example, the word "know" has three different representations over three different emotion categories. One possible reason can be that the preceding/following acoustic cue provides context information for a given region. The other reason lies with the role of prosody. The "I know" segment from both *sad*, Fig. 1c, and *happy*, Fig. 1d, categories has different representations, suggesting a strong relationship between the word importance and prosody for deciding emotion category. The attention on the phone "ay" across different segments and emotions shows the prosodic variation of the phone can constitute to different emotions.

6. Conclusions

Two contrasting systems are presented and evaluated on the commonly used IEM4 dataset, which contains elicited emotions. In this research, the contribution is two-fold. Primarily, in this work, a novel empirical bridge between the cognitive, phonetic theories and the computational models have been demonstrated by interpreting the deep neural models over acoustic speech emotion data. The attention vectors are interpreted by mapping them in the plots with the phones and sentences. Secondly, the relevance of acoustic context information is investigated, and it has been shown that even smaller acoustic cues hold emotion information. The paper also argues about the way speech emotion segments are labelled across long segments. The exact temporal limitations of this phenomenon are not clear. Future research is necessary to investigate that.

7. Acknowledgements

We sincerely thank Voicebase Inc. for their research funding and support.

8. References

[1] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. ICSLP*, 1996.

[2] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[3] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *International Journal of Speech Technology*, 2012.

[4] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, 2011.

[5] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artificial Intelligence Review*, 2015.

[6] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, 2006.

[7] R. Plutchik, "Emotion: Theory, research, and experience: Vol. 1. theories of emotion," *New York: Academic*, 1997.

[8] B. N. E. C. Y Susanto, A Livingstone, "The hourglass model revisited," in *IEEE Intelligent Systems*, 2020.

[9] P. Ladefoged and D. Broadbent, "Information conveyed by vowels," *Journal of the Acoustical Society of America*, 1957.

[10] P. Ladefoged, *Vowels and consonants*. Blackwell Oxford, UK, 2005.

[11] R. A. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey, "The contribution of consonants versus vowels to word recognition in fluent speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1996.

[12] D. Fogerty and D. Kewley-Port, "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," *The Journal of the Acoustical Society of America*, 2009.

[13] 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*, 2010.

[14] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, 2016.

[15] T. L. Nwe, S. W. Foo, and L. C. D. Silva], "Speech emotion recognition using hidden markov models," *Speech Communication*, 2003.

[16] Tin Lay Nwe, Foo Say Wei, and L. C. De Silva, "Speech based emotion classification," in *IEEE Region 10 International Conference on Electrical and Electronic Technology*, 2001.

[17] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[18] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech and Language*, 2015.

[19] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013.

[20] S. Zhang, T. Huang, and W. Gao, "Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition," in *Proceedings of the ACM on International Conference on Multimedia Retrieval (ICMR)*, 2016.

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

[22] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014.

[23] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Deep temporal models using identity skip-connections for speech emotion recognition," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017.

[24] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu, and H. Meng, "Speech emotion recognition using capsule networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[25] M. A. Jalal, E. Loweimi, R. K. Moore, and T. Hain, "Learning temporal clusters using capsule routing for speech emotion recognition," *INTERSPEECH*, 2019.

[26] R. Milner, M. A. Jalal, R. W. M. Ng, and T. Hain, "A cross-corpus study on speech emotion recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[27] M. A. Jalal, R. K. Moore, and T. Hain, "Spatio-temporal context modelling for speech emotion classification," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[28] D. Kewley-Port, T. Z. Burkle, and J. H. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *The Journal of the Acoustical Society of America*, 2007.

[29] M. J. Owren and G. C. Cardillo, "The relative roles of vowels and consonants in discriminating talker identity versus word meaning," *The Journal of the Acoustical Society of America*, 2006.

[30] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, 1999.

[31] F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman, "Some experiments on the perception of synthetic speech sounds," *The Journal of the Acoustical Society of America*, 1952.

[32] J. L. Miller, "On the internal structure of phonetic categories: A progress report," *Cognition*, 1994.

[33] T. Waaramaa, A.-M. Laukkanen, M. Airas, and P. Alku, "Perception of emotional valences and activity levels from vowel segments of continuous speech," *Journal of Voice*, 2010.

[34] A. Graves, N. Jaitly, and A. rahman Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013.

[35] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing*, 1997.

[36] R. B. et al., "Multi-modal sequence fusion via recursive attention for emotion recognition," in *Proc. ACL*, 2018.

[37] M. A. Jalal, R. K. Moore, and T. Hain, "Spatio-temporal context modelling for speech emotion classification," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[38] C. B. et al., "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, 2008.

[39] A. P. et al., "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[41] B. Desplanques and K. Demuynck, "Cross-lingual speech emotion recognition through factor analysis," in *INTERSPEECH*, 2018.

[42] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *INTERSPEECH*, 2017.

[43] P. Li, Y. Song, I. V. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *INTERSPEECH*, 2018.

[44] Z. Lian, J. Tao, B. Liu, and J. Huang, "Conversational emotion analysis via attention mechanisms," in *INTERSPEECH*, 2019.