SPEECH EMOTION RECOGNITION FROM GLOTTAL FLOW SIGNALS USING 1D CONVOLUTIONAL NEURAL NETWORKS

Itay Ben-Dom, Catherine I. Watson, Clare M. McCann

The University of Auckland, New Zealand
iben350@aucklanduni.ac.nz, c.watson@auckland.ac.nz, c.mccann@auckland.ac.nz

ABSTRACT

Speech emotion recognition can enhance machine-human interaction. While traditional speech emotion recognition systems tend to use prosodic features for analysis, recent studies show glottal features can provide valuable insight into distinguishing different types of emotional expressions. This paper presents a preliminary investigation into the effectiveness of employing the glottal waveform for speech emotion recognition tasks. The JL corpus, a strictly-guided, simulated emotional speech corpus in New Zealand English is used for this study. The modelling and classification tasks are conducted using 1D convolutional neural network. The results show that glottal signal enhances emotion classification accuracy compared to raw audio signals, even with data augmentation. This suggests the glottal source can serve as better alternative input to speech emotion recognition models, particularly when limited training data is available.

Keywords: speech emotion recognition, raw audio, glottal flow, IAIF, convolutional neural networks.

1. INTRODUCTION

The human voice is a powerful conveyor of information. It can provide cues about the speaker's biological, psychological, and emotional state [1]. The design of automatic speech emotion recognition (SER) systems has important applications in human-machine communication [2]. It is desirable for machines to be able to interpret the emotional state of the user and respond to it in an appropriate manner. The development of SER systems is a challenging task, however, due to difficulty in finding proper representations for emotion embedding in speech.

The emotional state of a speaker has direct effect on voice production, leading to physiological changes in respiration, phonation, and articulation [3]. Acoustic descriptors of emotional voice have traditionally been provided in terms of prosodic (e.g. pitch, speech rate) or spectral (e.g. MFCC) characteristics for speech analysis [4]. Over the past decade, however, there has been a growing body of evidence showing glottal source contains emotional content and glottal features can provide valuable insight into distinguishing different types of emotional expressions [5].

The source-filter theory describes human voice production as a linear time-invariant system [6]. During the process of speech production, air flow from the lungs passes through the glottis and generates a quasi-periodic signal called the glottal volume velocity waveform. Separating the source (glottal waveform) and filter (vocal tract) components enables the modeling of their distinct contributions. The glottal waveform can be separated from the speech signal via source-filter deconvolution. One popular approach to achieve this is through iterative adaptive inverse filtering (IAIF) [7]. IAIF has been used and evaluated in various experiments and it has been shown to yield rather robust estimates of the glottal flow.

In recent years, speech emotion recognition has been revolutionised by neural network models, following their demonstrated success in computer vision and speech recognition tasks [8]. Neural networks are a powerful tool and been shown to outperform traditional SER approaches. Convolutional neural networks (CNNs) are commonly used typologies in this area. One-dimensional CNNs that learn acoustic models directly from audio waveforms are becoming a popular method in audio processing due to the ability of these networks to take advantage of the signal's fine time structure [9]. Recent works explored the use of time-series signals as input for audio and speech processing with 1D CNNs [10]. The low-level layer of the model was shown to be able to learn frequency-selective filters from raw audio inputs [11]. It has been shown that features learned directly from the audio waveform can match, or outperform, the performance accuracy of a model trained on hand-crafted features [10, 12] Abdoli *et al* [13] proposed an end-to-end 1D CNN for environmental sound classification from raw audio signals. Different input sizes were

evaluated and the proposed architecture was shown to outperform 2D CNN models.

This study presents presents a preliminary investigation into the effectiveness of incorporating the glottal excitation source for emotion recognition task from speech signal. We utilise the architecture proposed in [13] to compare the model's emotion classification effectiveness when using two different input types; raw audio and glottal flow.To the best of our knowledge, this is the first attempt to consolidate glottal analysis with convolutional neural networks.

2. METHODOLOGY

The glottal flow is estimated from raw audio waveforms using inverse filtering algorithm from the GVV Toolbox package in R [14]. The glottal waveform is computed pitch synchronously using the iterative adaptive inverse filtering algorithm [15]. It is a two-stage iteration process governed by the principles of linear predictive coding, where the glottal waveform is computed by subtracting the contributions of the vocal tract and radiation load from the speech signal for each analysis frame. While the signal obtained through inverse filtering may be only an approximate of the actual glottal waveform and potentially result in experimental bias, the IAIF algorithm was chosen because it is widely used in the literature [16]. Each waveform is downsampled to 16 kHz and split into multiple input frame lengths to be fed into the network architecture; 0.1s (1,600 samples), 0.5s (8,000 samples), 1s (16,000 samples), and 2s (32,000 samples). The frames are extracted with 50% overlap and zero-padded (if required).

The baseline CNN architecture [13] is constructed by stacking two local feature learning block (LFLB), two convolutional layers, and three fully connected layers. The LFLB is designed to extract emotional features from time-series data. Each LFLB consists of one convolutional layer, one batch normalisation (BN) [17] layer, one rectified linear activation unit (ReLU) [18] layer, and one max-pooling layer. The convolution layer learns by sliding filters across the entire spatial dimension of the input. In a CNN, the filters in early layers learn low-level features, while the filters of the deeper layers learn high-level features that resemble concepts. The result of the convolution is a feature map. The filter size, stride, and number of filters are provided in details in [13]. All convolutions and pooling operations are one-dimensional, i.e. only along the axis representing time. The LFLB block is followed by two convolutional layers, each with batch normalisation and ReLU layers. The output of the final convolutional layer is flattened and used as input to two stacked fully connected layers, each with ReLU activation. Due to data sparsity during training, overfitting may be encountered. A popular and easy to implement regularisation technique is dropout [19]. Dropout is used to mitigate the problem of overfitting in neural networks by preventing co-adaptation of features during training. Dropout at a rate of 0.5 is used between each dense layer. The final output layer of this architecture is a softmax classifier with 5 neurons (matching the number of classes). We note that the network depth is proportional to the input size, where the number of convolutional layer is adjusted for the input length, with deeper models used to process longer audio segments.

3. EXPERIMENT

The JL corpus is a New Zealand English simulated emotion speech corpus, developed for emotion classification and synthesis in human-robot interaction research [20]. It contains 5 primary emotions: *happy*, *angry*, *sad*, *neutral* and *excited*. The speech samples were collected from four speakers (two male and two female) who are professionally trained voice actors of New Zealand English. The speech material consists of 15 neutral sentences with equal distribution of English long vowels - /a:/, /o:/, /i:/, /u:/. During the recording session each speaker was asked to repeat the 15 sentences twice for every emotion. Two separate recording sessions were conducted on different days to account for speaker's possible psycho-physiological abnormality at the time of recording. The audio clips were originally sampled at a rate of 44.1 kHz, but for the purposes of this study were downsampled to 16 kHz. Overall, the JL corpus contains 4 (speakers) $\times$ 5 (primary emotions) $\times$ 2 (repetitions) $\times$ 2 (sessions) = 1200 primary emotion sentences.

The 1D CNN network was implemented in TensorFlow Python library v2.9.0. The models were trained on NVIDIA T4 graphical processing unit (GPU) with 16 GB memory. Each model was trained with a batch size of 16 for 1000 epochs with early stopping. An Adadelta optimiser with 0.001 learning rate and 0.95 decay rate was used [21]. Considering speaker independent recognition, leave-one-speaker-out (LOSO) cross-validation was carried out for the discrimination of emotions in the JL corpus. The use of LOSO cross-validation ensures that the models are not over-training to a

Table 1: Test accuracy (%) for different input frame length and input type, with and without data augmentation (DataAug).

Audio Length (s)	*Model Input*			
	Raw Audio		**Glottal Flow**	
		DataAug		*DataAug*
0.1	34.8	35.7	37.7	**39.8**
0.5	37.3	41.4	40.9	**44.7**
1	38.8	43.8	42.2	**45.8**
2	35.0	42.4	36.5	**43.0**

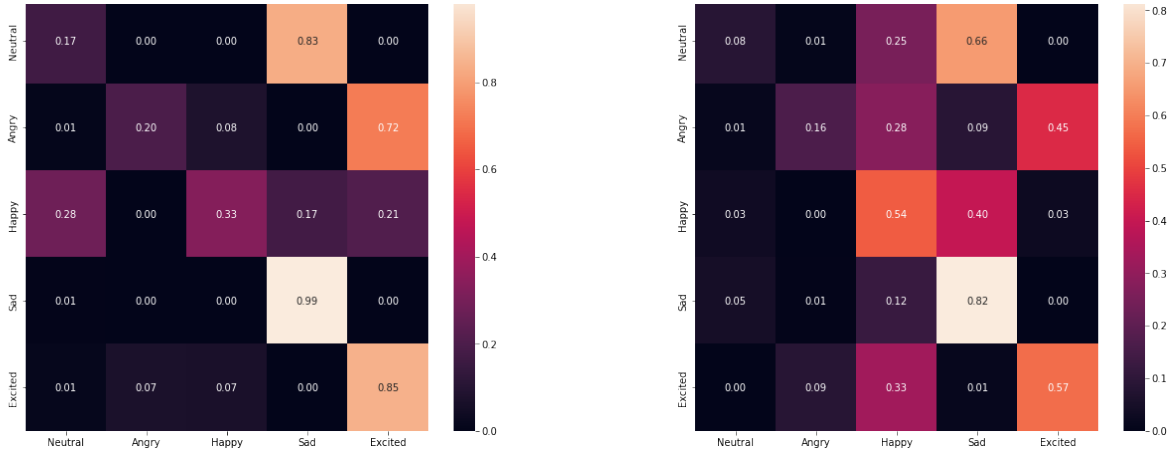
particular speaker. Since there are four speakers in the data set, the training set for each validation consists of three speaker, whereas the validation and test set were taken from an 80/20 split of the 4th speaker samples, respectively. Each validation was preformed 10-fold, with the final test accuracy calculated as the mean accuracy across all folds.

4. DISCUSSION

The models performances for each emotion category can be interpreted from the confusion matrix, shown in Figure 1. Figure 1a depicts the confusion matrix for best accuracy model with raw audio input. Figure 1b depicts the confusion matrix for best accuracy model with glottal waveform input. To study the model's performance for each emotion category, the confusion matrix was calculated by averaging the confusion matrix across ten cross-validation experiments for our best model architecture. For both input types, the high arousal emotions (e.g., angry, excited) are well separated from the low arousal ones (e.g., sad). Some of the most confused pairs are neutral-sad and angry-excited, which are different on the valence level. These erroneous classifications were extensively reported by other SER studies [22, 23]. The results show that classification from raw audio can discriminate between 2 emotions (sad, excited), while glottal waveform input showed discrimination between 3 emotions (happy, sad, and excited). These findings imply that a neural network trained on glottal waveforms is better equipped for discriminating between different emotions, as it has a greater capacity for emotion recognition.

The JL corpus presents speech samples for five primary emotions spoken in New Zealand English: *happy*, *angry*, *sad*, *neutral* and *excited*. Emotion classification performance was tested using both raw audio and glottal waveforms as inputs for an end-to-end 1D CNN models. The performance of the models was compared per input type and input length, and the results are presented in Table 1. The best performance was achieved using glottal input across all model topologies. It is apparent that models trained on glottal waveforms outperform models trained on raw audio. The 16,000-input (1 second) 1D CNN is preferable over other architectures, as it presents the best trade-off between the number of neural network parameters and prediction accuracy. This matches the findings of the original study [13].

The lack of large amounts of training data to train complex models is an ongoing challenge in speech emotion recognition. Insufficient training data can lead to overfitting and resulting in poor generalisation for unseen data. One common strategy to artificially expand the size of the training data set is data augmentation. Creating new samples through transformations of the existing data can help improve classification invariance and generalisation ability for neural network models. Growing the training set through data augmentation transformations have been shown to increased model accuracy in speech recognition tasks across all studies. General audio data augmentation techniques include tempo perturbation, loudness change, additive background noise, time-stretch, and pitch shift [24, 25]. One of the key requirements for data augmentation is to apply transformations to the inputs in a way that preserves their labels (ground truth does not change). In this study, three data augmentation techniques were applied on every utterance, doubling the size of the training set. Over-lapping windows were used to segment the input into various lengths. Since we are interested in continuous frame-based speech emotion recognition, 50% over-lap was applied between successive windows. This was done for both raw audio and glottal waveforms. A common technique for raw audio data augmentation is additive noise. For raw audio waveforms, additional transformation was carried out through distortion of the speech signal by adding noise from Laplacian distribution [26]. We propose a new data augmentation technique for glottal signals in the form of time-shift. We can shift the glottal signal left or right to create additional training data. We used a time-shift degree ranging from -5 to +5 ms. To the best of our knowledge this is the first attempt at applying data augmentation to glottal waveforms. The results show data augmentation boosted the performance accuracy by 3-7% for both raw audio and glottal waveforms, as shown in Table 1. These



(a) **(b)**

Figure 1: Normalised confusion matrix for best accuracy model with raw audio (a) and glottal waveform (b) as input sources.

findings support the current trend in literature and suggest data augmentation is a powerful tool for the training on models on smaller data sets.

Even with the application of data augmentation techniques, CNN models require substantial amount of training data to yield tangible results. The JL speech corpus used in this study has too few samples for the network to adequately capture the variety of emotions found in practice (e.g. *cold anger* vs *hot anger*). This is reflected in the emotion classification error rates observed in Figure 1. If the characteristics of the speech samples used for testing depart too much from those found in the training set, accuracy is expected to be low. When the training set is truly comprehensive, this problem is diminished. The small sample size of speakers (four) offers little variations for the CNN models to learn general characteristics and features. We noticed that the test accuracy for one of the male speakers was consistently lower than that of the other three speakers, which led to lower performance accuracy overall. In addition, since the corpus contains multiple repetitions of the same sentence (2 sessions with 2 repetitions), the corpus can be practically viewed as having only 600 unique samples, where the other 600 samples offer slight variation and therefore can be viewed as a form of data augmentation in itself. Thus, while the networks trained in the context of this work are not ready for practical use, the results they yielded provided a wealth of information that can be explored in future developments. Other investigations in literature often don't address this issue; they don't consider it significant or they use widely-used data sets. It is recommended that all results be interpreted in light of the limitations associated to the respective training data sets.

5. CONCLUSIONS AND FUTURE WORK

This paper demonstrates the effectiveness of using glottal waveforms as an alternative input for speech emotion recognition models. An end-to-end 1D CNN architecture was trained on both raw audio and glottal flow waveforms in order to discriminate between five primary emotions. The results show that neural network models trained on glottal inputs outperform the classification accuracy of those trained on raw audio. This reinforces the prevailing idea in the literature that the emotional content of speech signals is conveyed through the glottal pulse. Additionally, the efficacy of time-shift as a data augmentation approach for glottal signals has been demonstrated. Moreover, the influence of input dimensionality on classification performance underscores the significance of treating input dimensionality as a hyper-parameter in network architecture design. In conclusion, utilising the glottal flow as the input type for 1D CNNs shows potential for replacing raw audio, particularly in studies with limited data constraints. Building upon these findings, we intend to compare the performance accuracy of 1D and 2D CNN models for speech emotion recognition, using raw and spectrogram representations of the glottal flow, respectively. This will form the basis for our future research into speech emotion recognition in disordered speech.

6. REFERENCES

[1] A. Kappas, U. Hess, and K. R. Scherer, "Voice and emotion," in *Fundamentals of nonverbal behavior*, R. S. Feldman and B. Rime, Eds. Cambridge University Press, 1991, ch. 6, pp. 200–238.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] T. Johnstone, "The effect of emotion on voice production and speech acoustics," Ph.D. dissertation, The University of Western Australia, 2017, doi.org/10.31237/osf.io/qd6hz.

[4] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "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.

[5] L. He, M. Lech, J. Zhang, X. Ren, and L. Deng, "Study of wavelet packet energy entropy for emotion classification in speech and glottal signals," in *Fifth International Conference on Digital Image Processing (ICDIP 2013)*, Y. Wang and X. Yi, Eds., vol. 8878, International Society for Optics and Photonics. SPIE, 2013, p. 887834.

[6] G. Fant, *Acoustic theory of speech production*. Mouton, 1960.

[7] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2, pp. 109–118, 1992, eurospeech '91.

[8] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[9] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4624–4628.

[10] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," 2016.

[11] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6964–6968.

[12] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, "End-to-end speech recognition from the raw waveform," 2018.

[13] S. Abdoli, P. Cardinal, and A. Lameiras Koerich, "End-to-end environmental sound classification using a 1d convolutional neural network," *Expert Systems with Applications*, vol. 136, pp. 252–263, 2019.

[14] I. Ben-Dom, "Towards a voice analysis toolbox for the extraction, parametrisation and analysis of the glottal source waveform and its application for senescence voice," Master's thesis, The University of Auckland, 2017, http://hdl.handle.net/2292/35845.

[15] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2, pp. 109–118, 1992, eurospeech '91.

[16] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *INTERSPEECH*, 2016.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[18] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Gordon, D. Dunson, and M. DudÃk, Eds., vol. 15. Fort Lauderdale, FL, USA: PMLR, 11–13 Apr 2011, pp. 315–323.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[20] J. James, L. Tian, and C. I. Watson, "An open source emotional speech corpus for human robot interaction applications," in *INTERSPEECH*, 2018.

[21] M. D. Zeiler, "Adadelta: An adaptive learning rate method," 2012.

[22] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[23] L. Tian and C. Watson, "Emotion recognition using intrasegmental features of continuous speech," in *Proceedings of the 17th Australasian International Conference on Speech Science and Technology (SST2018)*, 2018, pp. 61–64.

[24] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[25] E. Lakomkin, M.-A. Zamani, C. Weber, S. Magg, and S. Wermter, "On the robustness of speech emotion recognition for human-robot interaction with deep neural networks," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 854–860, 2018.

[26] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.