Improving Emotional TTS with an Emotion Intensity Input from Unsupervised Extraction

Bastian Schnell[1,2], *Philip N. Garner*[1]

[1]Idiap Research Institute, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{bastian.schnell, phil.garner}@idiap.ch

Abstract

We aim to provide controls for emotion in synthetic speech. Many emotions are not displayed continuously in an otherwise emotional utterance; rather, the intensity varies with time. We show that an emotion recogniser is capable of producing a measure of emotion intensity via attention or saliency; this measure is appropriate to label utterances subsequently used to train a speech synthesiser. We evaluate novel and published means to do this showing that, whilst it is no longer state of the art for emotion recognition, attention is a good way to indicate emotion intensity for speech synthesis.

Index Terms: Emotional Speech Synthesis, TTS, Emotion Recognition, Saliency Mapping

1. Introduction

When text to speech synthesis (TTS) is used in a non-trivial application, it is desirable that the resulting synthetic speech conveys context-awareness using *affect*. For instance, in a speech to speech translation application, if the input speech (in L1) sounds, e.g., emphatic or angry, then the resulting speech (in L2) should convey the same qualities. In a dialogue application, the dialogue manager should be able to emphasise words that it wishes to clarify, and should respond to, say, frustration with empathy. Of course, this not only requires a suitably intelligent dialogue manager, but also a TTS system with controls for the appropriate affect variables. In this paper we are concerned with providing such controls for emotion.

While TTS systems have mastered human performance for neutral speech, emotional speech synthesis is still a challenge. For neutral speech large and high quality databases exist, but emotional databases are rare and mostly of low quality. It is certainly possible to record large amounts of a specific emotion and train the same systems as used for neutral speech. However, the range of emotions, varying intensities, the amount of languages, speaker variations, and the need to label each recording with the perceived emotion of multiple listeners makes recording alone a nearly infeasible task in terms of time and money. Modern emotional TTS research has identified three possible directions to solve these problems: 1) Increase the generalisability of the architectures on low data regimes; 2) increase the quantity of emotional data by voice or emotion conversion; and 3) increase the quality of the data. In the following we will highlight some recent work for each direction.

Databases with more expressive speech exist, especially audio books. Those databases cover a wider range of styles, but lack annotation of the expressed emotion or style. The lack of these annotations spawned a range of recent works focusing on increased model generalisability by utilising unsupervised methods to extract style embeddings from reference audio on a global [1, 2], clustered [3], or frame level [4, 5]. Some attempted controlling the expressiveness [1, 6]. However, controllability remains limited, especially for global embeddings.

Some work targets increasing the quantity of the expressive data. Huybrechts et al. [7] have used voice conversion to convert expressive recordings to the target speaker. In our recent work [8] we have converted neutral to emotional speech. The artificial data can then be used to train a TTS system.

We found limited work which attempts to increase the quality of the emotional data. Emotional databases usually have a single emotion label for every recording. We argue that this generalisation is misleading and that the emotion is localised within the utterance. This kind of annotation can lead to different emotion labels on words with lower emotional strength like conjunctions, while their acoustic features only marginally differ. Obviously, this impedes the learning of the model.

In this work we propose to add a frame-level emotion intensity to every sample, which is used as additional input to the TTS model. We present two methods to extract it from the recordings with pre-trained emotion recognisers. The simpler model contains a single attention layer, which allows use of the attention weights as emotion intensities. The other is a modern transformer model, where we exploit saliency maps to extract the intensity. The closest work to ours is that of Lei et al. [9]. They use relative attributes [10] to assign a level of emotional strength to each sample. In more recent work [11] they extended their method to phoneme level emotional strength.

We present our two methods for emotion intensity extraction as well as the method of attribute ranks of [11] in Section 2. We compare all three methods and a baseline without intensity input on the task of emotional TTS in Section 3. In this work we leave out the problem of generating the emotion intensity from text or extraction from a reference sample. Possible research directions to attack this problem are listed in the conclusions in Section 4.

2. Emotion intensity extraction

2.1. Attention LSTM

We use a simple emotion recogniser mostly resembling previous research [12] (Figure 1 left). It consists of a feature extraction part of 3 fully-connected layers with 256 neurons and a bidirectional LSTM (BiLSTM) with 128 neurons per direction. We apply dropout with a probability of 0.1 after each layer. Additionally, it contains an attention block with a single BiLSTM with 128 neurons per direction and a fully-connected layer without bias with a single output neuron. Its output respresents the unnormalised attention weights. As in previous work [12] we use a sigmoid activation, instead of the usual softmax, to obtain normalised attention weights. A sigmoid activation

ensures high activation levels over many frames, which leads to overall smoother attentions. This is especially desirable for our downstream task of emotional TTS. We use the predicted attention weights to compute a weighted sum over the outputs of the feature extraction part to create a single utterance level embedding of size 256. We pass this vector through a single fully-connected layer with as many neurons as emotion classes. All parameters are initialised using Xavier initialisation [13] with a uniform distribution, with one exception: The weights of the fully-connected layer with single output neuron in the attention block are initialised with samples from $\mathcal{N}(0, 0.1^2)$.

The openSMILE toolkit [14] is used to extract frame-level features (25 ms sliding window, 10ms shift). We use a 32-dim subset of the *IS09* feature subset composed of hand-crafted Low-Level Descriptors (pitch, energy, zero-crossing rate, voicing probability), 12 mel-frequency cepstral coefficients, and their first derivative. This subset is mean-variance normalised and forms the input to the emotion recogniser. To prevent overfitting we augment the input with random white noise with a standard deviation of 0.4. In contrast to previous research, this model accepts variable input lengths.

Training follows closely the procedure in previous work [12]. The model is trained with the Adam optimiser [15] ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1\text{E}{-8}$) with a learning rate of $3\text{E}{-5}$ on mini-batches of 32 for 200 epochs with the cross-entropy loss. To account for class-imbalance we weight the cross-entropy for each class c by a factor of $w_c = \frac{N_{tot}}{N_{classes} N_c}$, where N_{tot} is the total number of training utterances, $N_{classes}$ the number of different classes/emotions, and N_c the number of utterances of class c in the training set. All but the LSTM layers are regularised with l_2-regularisation with a factor of $5\text{E}{-2}$. We select the best model based on the summed Weighted Accuracy (WA) and Unweighted Accuracy (UA).

We argue that this emotion recogniser, once trained, will attend to the emotional parts of the utterance to make a decision. Thus it is reasonable to assume that the attention weights over an utterance give a good approximation of the emotion intensity.

2.2. Transformer

The above model is a very simple emotion recogniser and does not represent the state of the art. More complex architectures exist which do not allow a straight forward extraction of attention weights. In this section we investigate a more recent transformer model [16]. It consists of multiple self-attention blocks, which do not allow the extraction of attention weights in an obvious way. We make no claim that this model is the best emotion recogniser currently available; rather, we present a technique representative of more complex models without restrictions to their architecture to extract emotion intensities.

The transformer (Figure 1 right) consists of a feature extraction block with 4 fully-connected layers with 512 neurons and SeLU activation. Afterwards a positional encoding is added in form of a sinusoid with a large period. Dropout with 0.1 probability is applied on the latent features, which is then fed to two self-attention [17] blocks with 32 heads each. The resulting attention matrix is aggregated with five 2D convolutional layers with [30, 30, 30, 10, 6] output channels, a 5×5 kernel size, and a stride of 2×2. The flattened 936-dim output is projected with a fully-connected layer with 936 neurons and a final fully-connected output layer with as many neurons as emotion classes. After each but the last layer in the aggregation step, dropout with probability 0.2 and SeLU activation is applied. All parameters are initialised using Xavier initialisation

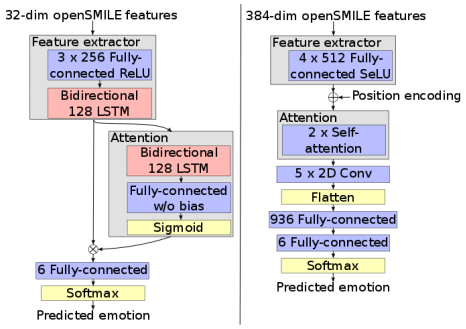

Figure 1: *Architectures of the emotion recognisers. Left: attention LSTM; right: transformer*

[13] with a uniform distribution.

As before we use the openSMILE toolkit to extract frame-level features (25ms window, 10ms shift). However, we use the entire 384-dim *IS09* features subset as input to the transformer model and we do not add any noise. The transformer model requires a fixed-length input. We use a sliding window of 500 frames with a step size of 50 frames previously found to perform best [16]. At inference time the final prediction is made by applying a softmax on the predicted classes of each window and averaging the results. Sequences are zero padded to match the window and step size, no frames are dropped.

During training we randomly select 500 frames from each input in the batch. We use the Adam optimiser ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1\text{E}{-8}$, no weight decay) with a learning rate of $1\text{E}{-5}$ for 170 epochs on a mini-batch size of 8 and PyTorch's *ReduceLROnPlateau* scheduler with default parameters.

To extract emotion intensities with the transformer model we propose to use saliency maps. Saliency maps are a common technique in vision-related machine learning tasks. They attempt to add interpretability to the neural network predictions. An increasing number of techniques exist with varying complexity [18, 19, 20, 21, 22]; we discuss some below. Saliency maps compute the importance of each input to the network's output, thus each openSMILE feature in each frame receives a value. To compute a scalar emotion intensity value we investigate the aggregation through *max* and *mean* operations. In the following we will give a high-level description of the techniques we use in our experiments (Section 3).

2.2.1. Saliency Maps

Input gradients [18] continues the backpropagation chain to the inputs and thus provides gradients of each input w.r.t. the correct class label. The idea is that the gradients indicate how much the class prediction is affected by a change in each input, thus representing its importance.

Since *input gradients* produces relatively noisy saliency maps **Smoothgrad** [19] attempts to smooth them over multiple observations. It achieves this by adding white noise to the input multiple times and computes the average input gradients for all iterations. The idea of *Smoothgrad* can be applied in many other saliency map techniques.

Input X Gradient [20] multiplies the *input gradients* with the input itself. The idea is that the gradient alone only indicates how important the feature is, but the input gives information on how strongly the feature is present. Together they provide a better abstraction of the feature importance.

Integrated Gradients [21] aggregates *input gradients* over

a linear interpolation between a baseline (the zero vector in our case) and the input. The idea is to capture *input gradients* that were steep at some of the interpolations but became flat for the input, as they are still important for the class prediction.

2.3. Attribute Rank

Recent work [9, 11] has used attribute ranks [10] to compute emotion intensities. We include this work as a competitive method here and give a brief overview. For data of two categories the ranking function computes the ranking/order of the data w.r.t. to a certain attribute, here emotion intensity. Once the ranking function is learned, it can assign an emotion intensity level to unseen emotional data. For completeness we give an example closely following that in [11].

We select all neutral N and happy H samples from the training set with acoustic features x_t with $t \in [1, \ldots, T]$ with $T = |N \cup H|$. We then form an ordered set O and an unordered set S of pairs. In the ordered set we pair an emotional sample of H with a neutral sample from N, indicating that the emotion intensity is higher in the samples of H than in those of N. In the unordered set we randomly create pairs of neutral-neutral and happy-happy samples, indicating that their rank should be similar. The goal is to learn a ranking function $r(x_t) = wx_t$ satisfying the following constraints as much as possible

$$\begin{aligned} \forall(i,j) \in O : wx_i > wx_j \\ \forall(i,j) \in S : wx_i = wx_j \end{aligned} \tag{1}$$

The problem can be relaxed with slack variables ξ_{ij} and γ_{ij} and solved by Newton's method.

In [11] a single openSMILE feature vector x_t is extracted for each utterance. Then the ranking function, i.e. the ranking vector w_m with $m \in [1, \ldots, M]$, is learned for each combination of neutral with the other M emotions. To obtain phoneme-level rankings openSMILE features are extracted for the segments corresponding to each phoneme. This requires a forced-alignment step for which we use the Montreal Forced Aligner [23]. We use a Python port[1] of the original code[2] of [10] with the default parameters for the Newton algorithm.

3. Experiments

For our experiments we select the SAVEE database [24]. It is an audio-visual British English database with sentences from TIMIT phonetically-balanced for each emotion. For each emotion 3 common, 2 emotion-specific, and 10 generic sentences (different for each emotion) were taken. For neutral the 3 common and $2 * 6$ emotion-specific sentences were additionally recorded, giving 30 neutral sentences in total. 4 males acted in 7 different emotions (neutral, anger, disgust, fear, happiness, sadness, and surprise) resulting in a total of 480 utterances. The audio was recorded at 44.1 kHz and has higher quality compared to most emotional databases. We do not use the visual information of the database. To compensate for loudness differences in speaker 'KL' we use a loudness normalization technique to normalize all samples to an average root-mean squared value of $RMS = 0.1$ with $\tilde{x} = x * \sqrt{(T * RMS^2)/(\sum^T (x - x_{mean})^2)}$. We also found background noise to degrade performance in some of the recordings. To reduce the noise we use a single channel spectral enhancement scheme [25] to pre-process the entire database.

[1] https://github.com/chaitanya100100/Relative-Attributes-Zero-Shot-Learning

[2] https://www.cc.gatech.edu/ parikh/relative.html

3.1. Emotion Intensity

To train emotion recognisers, the SAVEE database is rather limited. Thus we include the IEMOCAP [26] database in all strategies for emotion intensity extraction. It splits into 5 dialogue sessions of acted and spontaneous emotions with 2 different professional actors each, totalling 10 speakers and approximately 12 hours of 48 KHz recordings. At least 3 fluent English speakers annotated the perceived emotion and the final emotion label was chosen based on majority vote. While still in the database we exclude samples where no majority label was found, additionally we exclude the 'disgusted' emotion from our experiments, as it is both very hard to express and very rare in the database. We apply the same loudness normalization and noise reduction techniques as on SAVEE.

3.1.1. Emotion Recogniser

We train the attention LSTM (Section 2.1) and transformer (Section 2.2) emotion recogniser models on IEMOCAP with the parameters and inputs as defined in their respective section, using a random split of the 5th session for the validation and test set. We then fine-tune the models on SAVEE with the same parameters for 200 epochs and select the best model based on combined WA and UA on the validation set. For each emotion we select emotion specific utterances as test and validation set. Namely we use the 4th and 5th id as test set and the 6th and 7th id as validation set. We select the same ids for all speakers so that the content is unseen during training. Table 1 shows the metrics of the trained models on IEMOCAP and SAVEE. With the trained models we extract the emotion intensity. For the attention LSTM model these are simply the attention weights.

Table 1: *Weighted (WA) and Unweighted Accuracy (UA) of the emotion recogniser models after pre-training on IEMOCAP and fine-tuning on SAVEE excluding the disgusted emotion class.*

	IEMOCAP		**SAVEE**	
	WA	UA	WA	UA
Attention LSTM	54.7	40.3	62.5	60.4
Transformer	51.2	43.1	69.6	67.7

Table 2: *MSE between saliency maps and attention weights extraction on the attention LSTM model on SAVEE. Saliency maps abbreviated as IG: Input Gradient, Sg: Smoothgrad, IxG: Input x Gradient, IntG: Integrated Gradients*

Aggr.	**Smoothed**	**IG**	**Sg**	**IxG**	**IntG**
mean	no	1.46	1.451	1.626	1.62
mean	yes	0.625	**0.621**	1.179	1.312
max	no	1.466	1.461	1.5	1.56
max	yes	0.665	0.664	0.835	0.984

For the transformer model the variety of saliency maps (Section 2.2.1) allows multiple intensity curves (Figure 2). We extract emotion intensity using Input Gradients, Smoothgrad, Input X Gradient, and Integrated Gradients with *max* and *mean* aggregation. As the saliency maps can be noisy, we also experiment with smoothed versions obtained by a simple convolution with an 11 frames wide Hanning window (Figure 3). With informal listening we cannot select a best system. However, we find that the intensity weights extracted with the attention LSTM

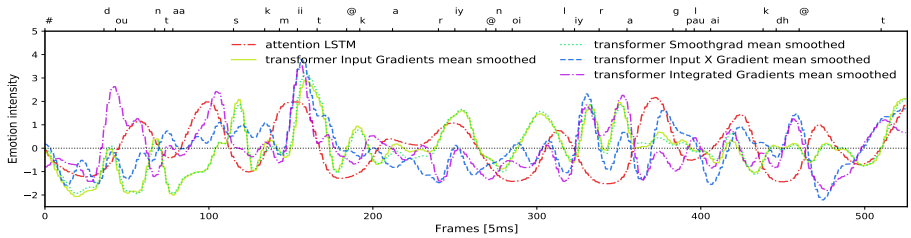

Figure 2: *Emotion intensities extracted with the attention LSTM model and different smoothed saliency maps for an angry utterance of speaker JK. For better comparison each intensity is mean-variance normalised based on its own statistics. The content is: "Don't ask me to carry an oily rag like that."*

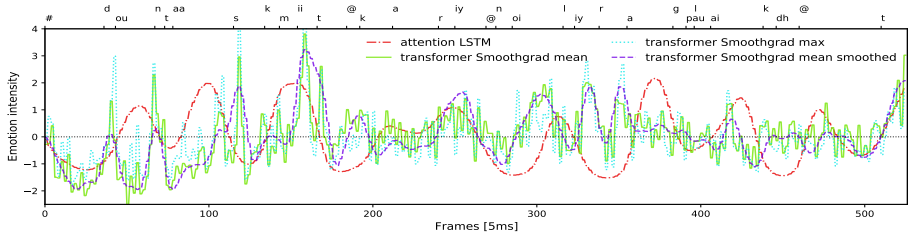

Figure 3: *Emotion intensities extracted with the attention LSTM model and with the Smoothgrad saliency map with max and mean aggregation as well as smoothed mean for the same utterance as in Figure 2*

model consistently produce more expressive speech than those extracted with saliency. Thus it is reasonable to interpret the saliency map as an approximation of the attention weights and select the saliency map which is closest to them. For that reason we extract the saliency maps on the attention LSTM model and compare them to the attention weights in terms of Mean-Squared-Error (MSE). As can be seen in Table 2 the closest saliency map is smoothgrad with smoothed *mean* aggregation.

3.1.2. Attribute Rank

While it is possible to learn the ranking just on the SAVEE database, we also include the IEMOCAP database for a fair comparison. Indeed, we found that rankings extracted on both databases outperform those learned only on SAVEE in informal listening tests. We exclude the SAVEE samples later used for validation/test set of the emotional TTS model (Section 3.2) when learning the ranking function. To form the unordered set we randomly form pairs for each sample in the neutral set of SAVEE. We then fill up the set with pairs from IEMOCAP (speaker independent selection) to reach 150 pairs. We perform the same with the respective emotion to obtain an unoredered set with 300 pairs. For the ordered set we randomly select a neutral SAVEE sample for each emotional SAVEE sample and again use IEMOCAP to fill up to 300 pairs. This procedure follows the one in [11].[3] With the learned ranking function we compute phoneme-level rankings for all SAVEE samples.

3.2. Emotional TTS

Our goal is to train an emotional TTS system with emotion intensity input on the SAVEE database. Due to the small size of SAVEE we cannot train a modern encoder-decoder network on it, as it quickly overfits before adapting the new speaking styles. Instead we rely on a classical RNN-based network, which has also been used in recent studies on emotional speech synthesis [27]. We use oracle durations in all our experiments, because duration prediction for emotional speech is a challenging problem on its own. The model consists of 2 fully-connected layers with ReLU activation and 1024 neurons, 3 BiLSTM layers with 512 neurons, and the final 97 dimensional output layer. 5% dropout is applied in all but the final layer. A 128-dim speaker and 64-dim emotion embedding is concatenated to the input of each layer. Additionally, we concatenate the mean-variance normalised emotion intensity input in all layers, which gives better results than concatenating it only to the input. For all neutral samples we set the emotion intensity to zero, indicating that there is no emotion present. We do not predict the emotion intensity internally, because we want to keep it as a tunable input.

The inputs to the model are 425 text-derived binary and numerical features normalised to $[0.01, 0.99]$, which were derived from the forced-aligned (with HTK [28]) phoneme sequence previously extracted with Festival [29]. The model predicts mean-variance normalised WORLD vocoder [30] features, consisting of linearly-interpolated log F_0 , a voiced/unvoiced flag, 30-dimensional mel-generalised cepstrum, and one Band Aperiodicity at 5 ms frame step, with their delta and double delta derivatives. The output is smoothed with the MLPG algorithm [31]. The WORLD vocoder is used to generate the waveform.

Even for our model the SAVEE database is too small to train a TTS system, so we instead pre-train on the WSJCAM0 database [32]. It is a large British English database with 92 speakers with 90 utterances each recorded at 16 kHz. We use only the head-mounted close-talking microphone recordings. We apply the same loudness normalisation and noise reduction techniques as on IEMOCAP and SAVEE (Section 3). The model is pre-trained for 35 epochs with a batch size of 16 and a learning rate of 0.001 and early stopping. We reduce the learning rate by a factor of 0.1 on validation loss plateaus. The adaptation to SAVEE is split into adaptation to the neutral subset of SAVEE first, and the entire database second. Each step is further divided into three phases. In the first phase only the speaker embedding is trained (10 epochs, lr=0.001), in the second phase

[3] We thank Shan Yang for the detailed description of the process.

the whole model is trained (10 epochs, lr=0.001), the last phase applies fine-tuning by repeating phase two with a smaller learning rate (10 epochs, lr=0.0001). The batch size in all phases is 16. In each phase early stopping is used and the best model is selected to continue with the next phase.

3.3. Subjective Results

In the subjective listening test we investigate how the TTS models compare in terms of perceived emotion and whether the audio quality is impacted. For the test we include five systems:

- **baseline**: TTS model without emotion intensity input
- **attention**: Attention weights from the attention LSTM
- **transformer**: Smoothgrad saliency map with mean aggregation and smoothing extracted with the transformer
- **rank**: Phoneme-level rankings extracted with the competitive technique [11]
- **ref**: Copy synthesis of the recordings

The test set consists of the same two utterances recorded for every emotion (7, including neutral, excluding disgusted) and every speaker (4 males). This makes 56 samples for each system. As we do not yet have a method to predict emotion intensity from text, we use the emotion intensity extracted from the reference audio by the respective technique. This gives an upper bound on the quality achievable with an emotion intensity input assuming that the prediction is perfect. We find that the emotion intensity input does not increase the expressiveness of the speech much. However, it offers an unprecedented control to tune the emotion intensity. Informal listening shows a greatly increased expressiveness, while still sounding natural, when scaling the input with a factor of **7**. The models have learned to connect certain speech properties with the intensity input, which allows scaling them in a natural way. In general higher intensities result in higher energy in the speech, which is desirable for all but the sad emotion. Thus all our tests use the scaled version except sadness.

36 listeners rated 25 randomly selected samples each in a 5-scale MOS test with 0.5 steps and also selected the emotion they perceived. Table 3 summarises the results. The *total* column includes the correct ratings on neutral. As many subtle emotions like fearful or surprised are rated as neutral, this number is biased. The *emo* column indicates the accuracy on the emotional samples only. On both metric the attention weight extracted with the attention LSTM model outperforms the other systems. It shows that an emotion intensity input increases the expressiveness of the speech, which is also perceivable by listeners. The *happy* emotion is almost never perceived. The low recognition rate of the reference samples indicates that it was not acted well enough. Providing a neutral reference during the listening test might facilitate its prediction.

It also shows that the quality of the emotion intensity matters as the phoneme level rankings perform much worse, this might be due to the phoneme-level granularity. The key benefit of the ranking function is that it requires very little training data. It might perform best when we do not include any IEMOCAP data. It outperforms the baseline system in a similar manner to that reported in the related work [11].

Interestingly, the saliency map extracted from the transformer model performs worse then the simple attention weight, even though the model is much more complex and achieves higher emotion recognition scores. All the saliency map techniques are developed for the field of vision, focusing on convolutional layers. A different type of saliency map might be necessary for speech tasks or more convolutional networks might allow better saliency maps. The benefit of the transformer model is that it will likely improve its emotion recognition performance with more training data compared to the attention LSTM model due to its small complexity. However, as long as no proper saliency map technique exists, we are limited to models that allow straight-forward extraction of emotion intensity.

Table 3: *Results of the subjective evaluation of perceived emotions in percentage. 'total' includes the neutral samples. Accuracy for each emotion is shown as well labelled as n: neutral, a: angry, f: fearful, h: happy, sa: sad, su: surprised.*

System	total	emo	n	a	f	h	sa	su
baseline	25.3	17.6	**72**	28	15	3	33	12
attention	**35.5**	**28.9**	70	**54**	13	**6**	**55**	**21**
transformer	26.7	20.6	60	33	**25**	0	41	8
rank	25.3	19.0	69	30	14	**6**	40	8
ref	45.9	40.3	75	74	23	19	31	54

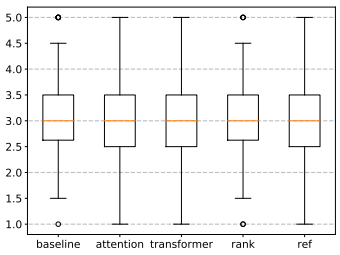

Figure 4: *Results of the 5-scale MOS test with 0.5 steps*

Figure 4 shows the results of the MOS test. None of the differences in the results are statstically significant in a two-tailed paired t-test with a p-value < 0.05. This includes the copy synthesis reference, which has other quality issues that were rated low by listeners. We can conclude that the proposed techniques do not deteriorate the audio quality.[4]

4. Conclusion and Future Work

We presented two techniques to extract an emotion intensity input from audio in an unsupervised way by utilising pre-trained emotion recognisers. We do not require emotion intensity labeling, but only emotion class labels. Thus one could also refer to it as weak supervision. From an emotion recognition network with a single attention layer we extract the attention weights as emotion intensity. From a transformer-based network we extract it using saliency maps. We show that the additional emotion intensity input improves an emotional TTS system; increasing the accuracy of which human listeners perceive the target emotion without degradation in signal quality. The simpler first method outperforms all others, including a recently published method for emotion intensity extraction by *relative attributes*.

For the tests we use oracle emotion intensity extracted from the reference. As the results show great improvements with an emotion intensity input, future research will focus on predicting it from text or conversion in speech-to-speech translation.

Ack.: This work was supported by the Swiss NSF grant number 185010: Neural Architectures for Speech Technology (NAST); http://p3.snf.ch/Project-185010

[4] Audio samples at www.idiap.ch/paper/ssw11_emotion_intensity/

5. References

[1] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[2] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[3] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," *Proc. Interspeech 2019*, pp. 4440–4444, 2019.

[4] S. Choi, S. Han, D. Kim, and S. Ha, "Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding," in *Proc. Interspeech 2020*, 2020, pp. 2007–2011. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2096

[5] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5911–5915.

[6] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7254–7258.

[7] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," *arXiv preprint arXiv:2011.05707*, 2020.

[8] B. Schnell, G. Huybrechts, B. Perz, T. Drugman, and J. Lorenzo-Trueba, "EmoCat: Language-agnostic emotional voice conversion," *arXiv preprint arXiv:2101.05695*, 2021.

[9] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 192–199.

[10] D. Parikh and K. Grauman, "Relative attributes," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 503–510.

[11] Y. Lei, S. Yang, and L. Xie, "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis," *arXiv preprint arXiv:2011.08477*, 2020.

[12] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 126–131.

[13] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[14] 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*, 2013, pp. 835–838.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[16] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition." in *Interspeech*, 2019, pp. 2578–2582.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[18] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[19] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[20] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[21] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3319–3328.

[22] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/80537a945c7aaa788ccfcdf1b99b5d8f-Paper.pdf

[23] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[24] S. Haq, P. J. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP08), Tangalooma, Australia*, 2008.

[25] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 61, 2015.

[26] 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, pp. 335–359, 2008.

[27] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Communication*, 2018.

[28] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using HTK." in *ICASSP (2)*, 1994, pp. 125–128.

[29] A. Black, P. Taylor, R. Caley, and R. Clark, "The festival speech synthesis system," 1998. [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival/

[30] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[31] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1315–1318.

[32] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, "WSJCAM0 corpus and recording description," *Cambridge University Engineering Department (CUED), Speech Group, Trumpington Street, Cambridge CB2 1PZ, UK, Tech. Rep. CUED/F-INFENG/TR*, vol. 192, 1994.