Interspeech 2018
2-6 September 2018, Hyderabad

Learning interpretable control dimensions for speech synthesis by using external data

Zack Hodari, Oliver Watts, Srikanth Ronanki, Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{zack.hodari, oliver.watts, srikanth.ronanki, simon.king}@ed.ac.uk

Abstract

There are many aspects of speech that we might want to control when creating text-to-speech (TTS) systems. We present a general method that enables control of arbitrary aspects of speech, which we demonstrate on the task of emotion control. Current TTS systems use supervised machine learning and are therefore heavily reliant on labelled data. If no labels are available for a desired control dimension, then creating interpretable control becomes challenging. We introduce a method that uses external, labelled data (i.e. not the original data used to train the acoustic model) to enable the control of dimensions that are not labelled in the original data. Adding interpretable control allows the voice to be manually controlled to produce more engaging speech, for applications such as audiobooks. We evaluate our method using a listening test.

Index Terms: controllable speech synthesis, expressive speech synthesis, emotion recognition

1. Introduction

Typical text-to-speech (TTS) voices do not allow arbitrary aspects of speech to be controlled, even though such control would be useful in many applications. Supervised methods for style control (such as i-vectors [1] and LHUC [2, 3]) require data which contains variation along the dimensions to be controlled, this variation must be labelled. Annotation of much of the variation in natural speech is an unsolved problem; unannotated and, therefore, uncontrolled variation in training data is detrimental to TTS voices. TTS datasets are typically designed to exclude variation that we might wish to control, in pursuit of more consistent speech. 'Found' data which has been produced with genuine communicative intent – and which might therefore contain interesting variation – is generally not labelled in the way that is needed for learning control. Although this kind of naturally variable and labelled data is possible to create, doing so is prohibitively expensive and time-consuming.

Methods for learning dimensions of latent variation in speech data have been proposed [4, 5], however, by their unsupervised nature, the control vectors learnt are often not interpretable by human users. Voices resulting from these unsupervised methods might be suitable for use in dialogue systems where control can be automated. However, for applications such as the semi-automatic production of audiobooks it may be desirable for human curators to directly control the synthetic speech, this requires the control vectors associated with the voice to be interpretable.

In this paper, we propose an alternative approach which addresses many of these problems. The approach uses external labelled data as a means to label our synthesis dataset; the only constraint on the synthesis data itself is that it must include sufficient variation. We demonstrate the method on the task of emotion control, which is challenging because it is realised through complex acoustic changes [6]. We demonstrate the ability of our method to adapt a deep neural network (DNN) statistical parametric speech synthesis (SPSS) system and show that the use of external labelled data enables style adaptation.

Methods for speaker adaptation (which could just as well be applied to the adaptation of style) fall into three broad categories: feature-space normalisation, model-based adaptation, and auxiliary features. Feature-space normalisation aims to perform speaker-dependent normalisation on the acoustic parameters; for example, the linear input network (LIN) [7] adds a speaker-dependent layer to normalise the inputs into a speaker-independent space. Model-based adaptation adjusts the model in order to handle different speakers (e.g. LHUC [2]). In the auxiliary feature approach, the model learns to adapt with respect to some additional input features (such as speaker-dependent i-vectors [1]).

Schroder et al. [8] review several techniques for expressive speech synthesis, including methods which can be applied to formant synthesis, diphone synthesis, unit-selection synthesis, and HMM-based synthesis. Style-dependent models for emotive speech synthesis based on unit-selection and HMMs were investigated by Barra-Chicote et al. [9], i.e. separate models were trained for each emotion. This multiple model technique is useful when there are distinct classes and sufficient data is available for each class, but as this is often not the case, adaptation of a single model is an attractive option.

Adaptation for the purpose of emotion control in statistical parametric speech synthesis (SPSS) was first demonstrated by Yamagishi et al. using hidden Markov models (HMMs) [10]. A more complex form of adaptation, cluster adaptive training (CAT), was demonstrated for expressive synthesis [11]. CAT is useful for training on diverse data, and can be regarded as a form of model combination which jointly learns the parameters of multiple models and how to combine them. While CAT has been demonstrated for DNN acoustic models in speech recognition [12], complex model combination may be unnecessary for neural networks which are inherently able to model diverse data. As simpler methods such as input codes (i.e. auxiliary features) have shown good performance for the adaptation of speaker, gender, and age [13], we use auxiliary features to perform adaptation.

2. Controllable SPSS

We achieve controllable SPSS through style adaptation, making use of our control vector (emotion labels) as an auxiliary input feature: this simple method works reasonably well for DNN-based models [3, 13]. For training, we employ an emotion recognition model (previously trained on the external data) to label the synthesis training data, as described in Section 2.1. For synthesis, there are two available techniques to derive the control vectors: when a natural rendition of the required sen-

10.21437/Interspeech.2018-2075

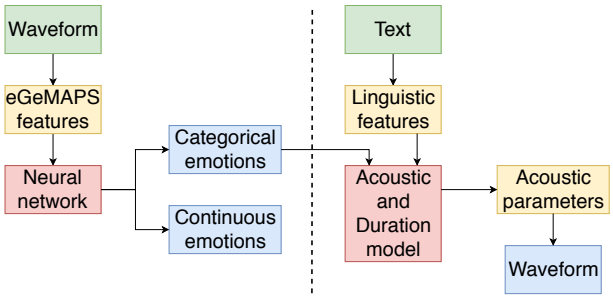

Figure 1: *Controllable SPSS system using external data. In the left half of the figure, an emotion recognition model is trained on external data. In the right half, the predicted labels are used as auxiliary features in a DNN-based speech synthesiser. Green boxes indicate inputs, yellow boxes indicate intermediate data representations, red boxes indicate models, and blue boxes indicate outputs.*

tence exists, we can extract the oracle control vector using the emotion recognition model, as in training. Or, if we are generating novel sentences from text alone, the control vector can be manually specified, as discussed in Section 2.2.

The control vector used in our experiments is a 4-dimensional probability vector, where each control dimension corresponds to a categorical emotion class with which the external dataset is already labelled (Section 3.1.1). Deriving the control vector from human annotations should result in interpretable control. We also experimented with appraisal-based (i.e. continuous) emotion labels as the control dimensions since these were also available in the external dataset. The recognition model can be trained to predict whatever type of label is present in the external dataset. However, the labels being predicted must describe an aspect of speech that varies in the synthesis dataset itself. The labels placed on the synthesis dataset are then used as input to the duration and acoustic models (Figure 1), concatenated with the linguistic features – created from text in the usual way, in our case using Festival's front end [14]. The acoustic parameters are extracted using GlottHMM (for F_0) [15] and STRAIGHT [16] (for spectrum and aperiodicity).

2.1. Emotion recognition model

The emotion recognition model is a simple feed-forward neural network trained using multi-task learning [17] to predict two types of emotion labels – categorical (e.g. happy) and continuous (e.g. level of arousal) labels. For each task, there is a private hidden layer just before the output layers, not shared between the two tasks.

The input is the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [18], a feature vector designed to be predictive of emotion. The eGeMAPS features are motivated by their ability to model perceptually-relevant changes in speech, and have proven performance in empirical studies. Each of the 88 features are utterance-level functionals of low-level descriptors (LLDs). The LLDs include frame-level energy, spectral, cepstral, prosodic, and voicing descriptors.

2.2. Constructing control inputs for synthesis

For training, we label the synthesis data using predictions from the emotion recognition model. To synthesise new sentences (right half of Figure 1) we must construct a control vector in place of the prediction used during training. In our experiment, they were provided, sentence-by-sentence, by one of the authors using a simple graphical interface.

3. Experiments

3.1. Learning the control dimensions

We built our recognition model in TensorFlow [19] and the source code for this system is available online[1].

3.1.1. Emotion recognition database

For training and evaluation of our emotion recognition model, we use the *Interactive Emotional Dyadic Motion Capture* dataset (IEMOCAP) [20] which contains 12.5 hours of data from both scripted and improvised sessions between two actors. There are 5 male and 5 female actors; each mixed-gender dyad was recorded for two sessions of about 1 hour. Sessions contain an average of 15 conversations, each designed to produce one of 5 categorical emotions – anger, sadness, happiness, frustration, and neutral. For improvised conversations, the actors were given hypothetical scenarios designed to elicit a specific emotion.

Each utterance (average length 4.5 seconds) was labelled by 3 annotators for both categorical and continuous emotion – arousal, valence, and dominance. After annotation, the authors added disgust, fear, excitement surprise, and "other", for a total of 10 categories. However, we used a subset of the data that contains only happy, sad, angry, and neutral utterances.

3.1.2. Classification results

We split the data into 5 cross-validation folds, taking 4 dyads for the training set in each fold. We split our test set (1 dyad per fold) in half; similar to Lee et al. [21], we use 1 speaker for parameter tuning, and the other for held-out evaluation. The results presented use these held-out speakers[2].

Our best system achieved an emotion classification accuracy of 62.9% and employed a neural network with one shared layer of 200 units and 2 private layers of 20 units each. We tried many architectures but found that performance ceilings at around 62% regardless. Table 1 reports comparable results from the literature, all using IEMOCAP and predicting the same 4 emotions. Our simple architecture produces acceptable results.

In the next section we describe how we train a TTS system on data lacking categorical emotion labels, by labelling it with the above IEMOCAP-trained model. The categorical and continuous emotion predictions for our TTS dataset are presented in Figures 2 and 3.

3.2. Controllable SPSS

3.2.1. TTS database

To train an expressive voice using SPSS we require a dataset containing sufficiently varied speech; we use the Blizzard Challenge 2017 dataset [28] provided by Usborne Publishing.

This dataset contains 6.5 hours of professionally-recorded audiobook data from a female speaker of Standard Southern British English. The stories are mostly aimed at 4–6 year olds, and include: traditional stories (e.g. *Little Red Riding Hood*); simplified Shakespeare (e.g. *Macbeth*); and non-fiction (e.g.

[1] https://github.com/ZackHodari/IS18_control_space

[2] The held-out speaker's gender was alternated in each fold, to avoid the test set consisting of a single gender.

Table 1: *Overview of comparable IEMOCAP recognition results, classifying the same emotions; angry, happy, sad, and neutral.*

Method	**Input features**	**Accuracy**
feed-back long short-term memory (LSTM), attention[22]	predicts using previous emotion	60.8%
ensemble support vector machine[23]	12 MFCCs, jitter, shimmer	60.9%
convolutional neural network, multiple kernel learning[24]	ComParE 2016	61.3%
deep belief network, support vector machine[25]	ComParE 2010	62.5%
feed-forward neural network (this paper)	eGeMAPS	**62.9%**
recurrent neural network, extreme learning machine[21]	MFCCs, F_0, voice probability, zero-crossing rate	63.9%
progressive deep neural networks[26]	eGeMAPS	65.7%
convolutional neural network, LSTM[27]	Spectrogram (cropped to 3 seconds)	68.8%

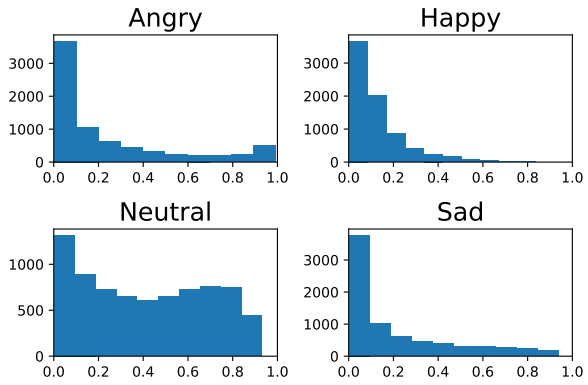

Figure 2: *Histograms of cross-corpus predictions on the Blizzard data for categorical emotion labels.*

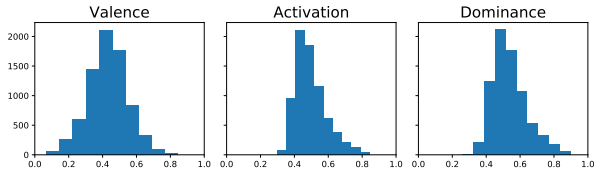

Figure 3: *Histograms of cross-corpus predictions on the Blizzard data for appraisal-based (i.e. continuous) emotion labels.*

The Story of Chocolate). Many of the stories include substantial amounts of direct speech. Stories are read expressively and the voice actor is consistent, making it appropriate for our work. We use the same training-validation-test split as Watts et al. [4].

3.2.2. System description

We created our systems using the open-source toolkit Merlin[3] [29]. Following the *fls_blizzard2017* recipe from Merlin, we trained two systems: a baseline model (*DNN-B*) and our proposed system (*DNN-C*) where control vectors are added as auxiliary features. Both systems were trained using feed-forward DNNs comprising 6x1024 `tanh` layers. The STRAIGHT vocoder [16] was used for waveform generation.

3.2.3. Objective results

In Table 2 we present results for systems *DNN-C* and *DNN-B* – with and without our control vector. The control vector is not expected to reduce objective error; these measures simply confirm that overall quality is about the same in both cases.

[3]https://github.com/CSTR-Edinburgh/merlin

Table 2: *Objective results of two SPSS voices with and without control vectors.*

	Objective metric			
	MCD (dB)	BAP (dB)	log F_0 (RMSE)	VUV (error %)
DNN-B (baseline)	**5.650**	**0.075**	51.209	**7.451**
DNN-C (with control)	5.719	0.076	**50.624**	7.551

3.2.4. F_0 variation

To demonstrate that the control vectors are able to produce meaningful variation during synthesis, Figure 4 shows F_0 for a single sentence, synthesised using 4 different one-hot control vectors, representing the 4 different emotions.

3.3. Subjective evaluation

We performed subjective experiments using three systems[4]: a baseline voice (*DNN-B*) using the standard SPSS recipe from Merlin, a randomly controlled voice (*DNN-R*), and our proposed system (*DNN-C*) where control vectors were added. *DNN-R* is the same as *DNN-C*, but with each element of the control vector set randomly between 0 and 1, sentence-by-sentence.

31 university students (26 female, 5 male) were paid to carry out the test[5] in sound-proof booths using using Beyerdynamic DT770 headphones. The test typically took an hour.

3.3.1. Simple emotion control

For this experiment, the test material comprised 50 audiobook sentences from the test set used in the Blizzard Challenge 2017 [28]. The control vector can be used to modify the perceived emotion and, as described in Section 2, the control vector comprises four values between 0 and 1. We prepared 4 versions of each test sentence using a one-hot vector (i.e. we used 'canonical' emotions) for each emotion in turn.

Participants performed a forced-choice labelling task, choosing the closest emotion out of the 4 classes for each presented utterance. Table 3 presents the results as a confusion matrix; the average classification accuracy is 41% (chance level is 25%).

To place this result in context, we analysed the inter-annotator accuracy of the IEMOCAP dataset, which is on average 48% for 10 emotion classes (where all speech is natural, of course). This is in line with the survey and evaluation by Banse et al. [6], who concluded that human performance at emotion

[4]Speech samples are available at https://zackhodari.github.io/IS18_control_space.html

[5]Implemented with BeaqleJS [30], available at https://github.com/HSU-ANT/beaqlejs

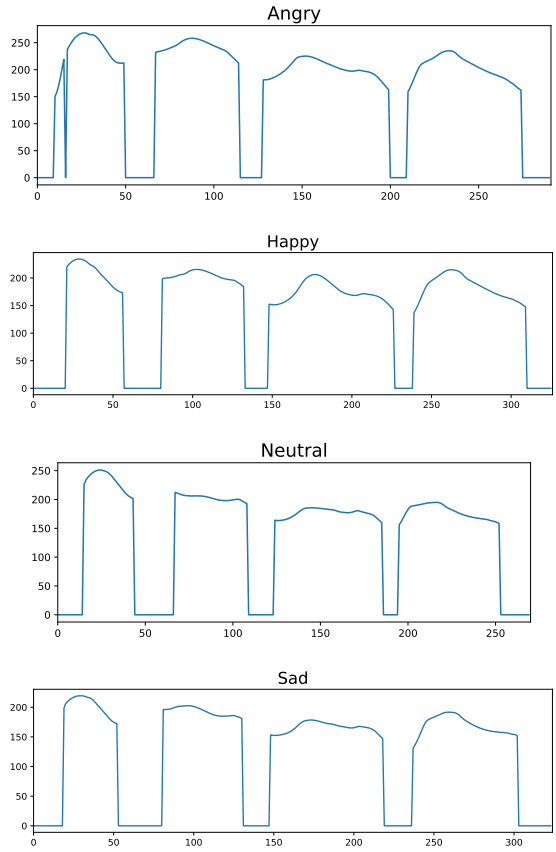

Figure 4: *Demonstration of F_0 variation as control vector is changed*

labelling of natural speech is around 50%. Banse et al. cite two other studies evaluating human performance on a 5-class emotion classification task, reporting accuracies of 64% [31] and 56% [32] respectively, again on natural speech. With this context, human performance of 41% on our synthetic speech is satisfactory. Our controllable system is able to modify perceived emotion.

The very low accuracy for 'happy' (13%), and the perception by listeners that 36% of 'happy' sentences sound 'angry' is likely a symptom of the imperfect labelling from our emotion recognition model. Predictions for the dimension corresponding to happy in the control vectors are heavily skewed towards zero in Figure 2, this may lead to compounding errors. While we did not investigate this thoroughly, the voice is able to perform adaptation successfully, and the dimensions of control are mostly interpretable by human listeners.

3.3.2. Creating variation when reading longer texts

It seems reasonable to think that listeners will prefer sentence-by-sentence variation when listening to an extended text (e.g., an audiobook), rather than the same speaking style for every sentence. In this experiment, we demonstrate that this variation should not be random, but must be appropriate to the text.

The test material was 17 short audiobook paragraphs, again from the Blizzard Challenge 2017 test set, with an average length of 24 seconds. Listeners were presented with pairs of versions of the same paragraph, generated using either systems *DNN-B* & *DNN-C* or systems *DNN-R* & *DNN-C*, and asked to "choose the paragraph you would prefer if you were listening

Table 3: *Confusion matrix for the forced-choice emotion classification task; the accuracy for each emotion is in bold face.*

Correct class	Predicted class: Angry	Happy	Neutral	Sad
Angry	**30%**	51%	13%	7%
Happy	36%	**13%**	29%	22%
Neutral	10%	15%	**66%**	10%
Sad	10%	4%	30%	**56%**
	MEAN ACCURACY **41%**			

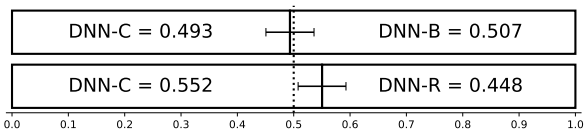

Figure 5: *Pairwise preference ratios including 95% confidence interval*

to an audiobook for pleasure". System *DNN-C* was controlled by one of the authors using the UI mentioned in Section 2.2, the trial-and-error process of finding a satisfactory control vector for each sentence took between 2 and 3 minutes per paragraph.

The ratios and 95% confidence intervals for the two pairwise tests are presented in Figure 5. Significance was calculated using the binomial confidence interval. There is no significant difference between *DNN-B* and *DNN-C*, so we can only conclude that our controlled system is at least as good as baseline. This is still a positive result because it is important that imposing control does not degrade the quality of the voice. *DNN-R* is significantly less preferred than *DNN-C*, showing that users only like variation when it fits the text at least to some extent.

4. Conclusions and future work

We have shown that by learning an emotion classifier from an external dataset, we are able to label a previously unlabelled synthesis dataset (from a different speaker) and from that create a controllable text-to-speech voice. The resulting control is interpretable to a human operator, and produces emotion variation that is perceivable by listeners. We also found that listeners significantly prefer appropriate variation over random variation when listening to audiobook paragraphs, although we have not yet created variation that is preferred over a non-varying baseline. The paragraphs used in the Blizzard Challenge 2017 are actually very short (this is necessitated by their listening test design and the large number of systems being compared). Perhaps there is simply not enough time for listeners to perceive sentence-to-sentence variation, and on a longer text the variation would be more noticeable.

Our method can be applied to other types of variation in speaking style. All that is needed is an existing dataset labelled with dimensions of interest. This dataset does not need to be recorded in high quality, nor does the speech need to be transcribed. Our method is able to transfer these labels to the single-speaker studio-quality synthesis data that is necessary for training a text-to-speech system.

Acknowledgements: This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

5. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 171–176.

[3] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Principles for learning controllable TTS from annotated and latent variation," in *Proc. Interspeech*, 2017, pp. 3956–3960.

[6] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[7] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[8] M. Schröder, "Expressive speech synthesis: Past, present, and possible futures," in *Affective information processing*. Springer, 2009, pp. 111–126.

[9] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech communication*, vol. 52, no. 5, pp. 394–404, 2010.

[10] J. Yamagishi, T. Masuko, and T. Kobayashi, "HMM-based expressive speech synthesis - Towards TTS with arbitrary speaking styles and emotions," in *Proc. of Special Workshop in Maui (SWIM)*, 2004.

[11] L. Chen, M. J. Gales, N. Braunschweiler, M. Akamine, and K. Knill, "Integrated expression prediction and speech synthesis from text," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 323–335, 2014.

[12] T. Tan, Y. Qian, and K. Yu, "Cluster adaptive training for deep neural network based acoustic model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 459–468, 2016.

[13] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4905–4909.

[14] P. Taylor, A. W. Black, and R. Caley, "The architecture of the Festival speech synthesis system," 1998.

[15] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based finnish text-to-speech system utilizing glottal inverse filtering," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[16] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[17] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[18] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[19] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[20] 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.

[21] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition." in *INTERSPEECH*, 2015, pp. 1537–1540.

[22] R. Zhang, A. Atsushi, S. Kobashikawa, and Y. Aono, "Interaction and transition model for speech emotion recognition in dialogue," *Proc. Interspeech 2017*, pp. 1094–1097, 2017.

[23] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of SVM trees for multimodal emotion recognition," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–4.

[24] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 439–448.

[25] R. Xia and Y. Liu, "Leveraging valence and activation information via multi-task learning for categorical emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5301–5305.

[26] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," *arXiv preprint arXiv:1706.03256*, 2017.

[27] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," *Proc. Interspeech 2017*, pp. 1089–1093, 2017.

[28] S. King, L. Wihlborg, and W. Guo, "The Blizzard challenge 2017," 2017.

[29] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[30] S. Kraft and U. Zölzer, "BeaqleJS: HTML5 and javascript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference, Karlsruhe, DE*, 2014.

[31] R. v. Bezooijen, "The characteristics and recognisability of vocal expression of emotions," 1984.

[32] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and emotion*, vol. 15, no. 2, pp. 123–148, 1991.