Emotion Recognition in Speech using Cross-Modal Transfer in the Wild

Samuel Albanie*, Arsha Nagrani*, Andrea Vedaldi, Andrew Zisserman
Visual Geometry Group, Department of Engineering Science, University of Oxford
{albanie,arsha,vedaldi,az}@robots.ox.ac.uk

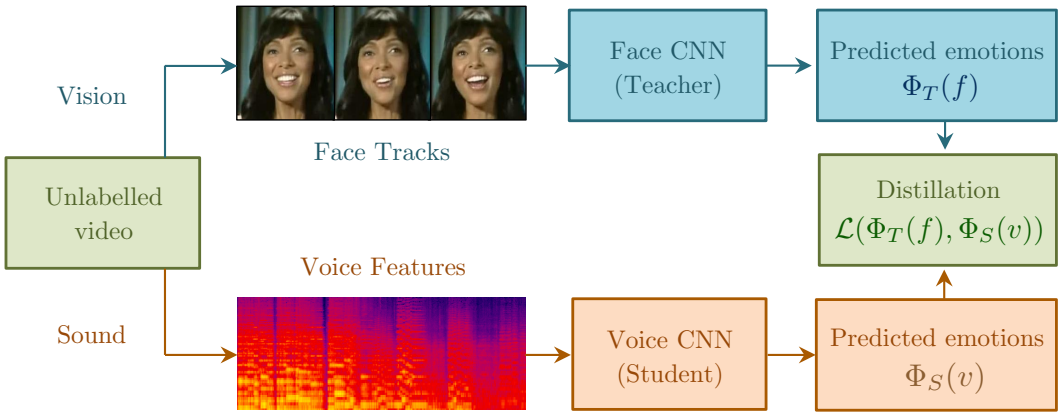

Figure 1: Cross-modal transfer: A CNN for speech emotion recognition (the student, Φ_S) is trained by distilling the knowledge of a pretrained facial emotion recognition network (the teacher, Φ_T) across unlabelled video. The student aims to exploit redundancy between the audio and visual signals v and f to learn embeddings, reducing dependence on labelled speech.

ABSTRACT

Obtaining large, human labelled speech datasets to train models for emotion recognition is a notoriously challenging task, hindered by annotation cost and label ambiguity. In this work, we consider the task of learning embeddings for speech classification without access to any form of labelled audio. We base our approach on a simple hypothesis: that the emotional content of speech correlates with the facial expression of the speaker. By exploiting this relationship, we show that annotations of expression can be transferred from the visual domain (faces) to the speech domain (voices) through *cross-modal distillation*. We make the following contributions: (i) we develop a strong teacher network for facial emotion recognition that achieves the state of the art on a standard benchmark; (ii) we use the teacher to train a student, *tabula rasa*, to learn representations (embeddings) for speech emotion recognition *without access to labelled audio data*; and (iii) we show that the speech emotion embedding can be used for speech emotion recognition on external benchmark datasets. Code, models and data are available[1].

[1]http://www.robots.ox.ac.uk/~vgg/research/cross-modal-emotions

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10.
https://doi.org/10.1145/3240508.3240578

KEYWORDS

Cross-modal transfer, speech emotion recognition

ACM Reference Format:
Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, Andrew Zisserman, Visual Geometry Group, Department of Engineering Science, University of Oxford, {albanie,arsha,vedaldi,az}@robots.ox.ac.uk. 2018. Emotion Recognition in Speech using Cross-Modal Transfer in the Wild. In 2018 ACM Multimedia Conference (MM'18), October 22–26, 2018, Seoul, Republic of Korea. ACM, Seoul, Korea, 10 pages. https://doi.org/10.1145/3240508.3240578

1 INTRODUCTION

Despite recent advances in the field of speech emotion recognition, learning representations for *natural* speech segments that can be used efficiently under noisy and unconstrained conditions still represents a significant challenge. Obtaining large, labelled human emotion datasets 'in the wild' is hindered by a number of difficulties. First, since labelling naturalistic speech segments is extremely expensive, most datasets consist of elicited or acted speech. Second, as a consequence of the subjective nature of emotions, labelled datasets often suffer from low human annotator agreement, as well as the use of varied labelling schemes (i.e., dimensional or categorical) which can require careful alignment [46]. Finally, cost and

*Equal contribution.

time prohibitions often result in datasets with low speaker diversity, making it difficult to avoid speaker adaptation. Fully supervised techniques trained on such datasets hence often demonstrate high accuracy for only intra-corpus data, with a natural propensity to overfit [42].

In light of these challenges, we pose the following question: is it possible to learn a representation for emotional speech content for natural speech, from *unlabelled* audio-visual speech data, simply by transferring knowledge from the facial expression of the speaker?

Given the recent emergence of large-scale video datasets of human speech, it is possible to obtain examples of unlabelled human emotional speech at massive scales. Moreover, although it is challenging to assess the accuracy of emotion recognition models precisely, recent progress in computer vision has nevertheless enabled deep networks to learn to map faces to emotional labels in a manner that consistently matches a pool of human annotators [1]. We show how to transfer this discriminative visual knowledge into an audio network using unlabelled video data as a bridge. Our method is based on a simple hypothesis: that the emotional content of speech correlates with the facial expression of the speaker.

Our work is motivated by the following four factors. First, we would like to learn from a large, *unlabelled* collection of 'talking faces' in videos as a source of free supervision, without the need for any manual annotation. Second, evidence suggests that this is a possible source of supervision that infants use as their visual and audio capabilities develop [30]. Newborns look longer at face-like stimuli and track them farther than non-face-like stimuli (Goren et al. [29]; Johnson et al. [38]), and combining these facial stimuli together with voices, detect information that later may allow for the discrimination and recognition of emotional expressions. Our third motivation is that we would like to be able to handle ambiguous emotions gracefully. To this end, we seek to depart from annotation that relies on a single categorical label per segment, but instead incorporate a measure of uncertainty into the labelling scheme, building on prior work by [66] and [32]. Finally, accepting that the relationship between facial and vocal emotion will be a noisy one, we would like to make use of the remarkable ability of CNNs to learn effectively in the presence of label noise when provided with large volumes of training data [45, 59].

We make the following contributions: (i) we develop a strong model for facial expression emotion recognition, achieving state of the art performance on the FERPlus benchmark (section 3.1), (ii) we use this computer vision model to label face emotions in the VoxCeleb [50] video dataset – this is a large-scale dataset of emotion-unlabelled speaking face-tracks obtained in the wild (section 4); (iii) we transfer supervision *across modalities* from faces to a speech, and then train a speech emotion recognition model using speaking face-tracks (section 5); and, (iv) we demonstrate that the resulting speech model is capable of classifying emotion on two external datasets (section 5.2). A by-product of our method is that we obtain emotion annotation for videos in the VoxCeleb dataset automatically using the facial expression model, which we release as the EmoVoxCeleb dataset.

2 RELATED WORK

Teacher-student methods. Teaching one model with another was popularised by [12] who trained a single model to match the performance of an ensemble, in the context of model compression. Effective supervision can be provided by the "teacher" in multiple ways: by training the "student" model to regress the pre-softmax logits [7], or by minimising cross entropy between both models' probabilistic outputs [43], often through a high-temperature softmax that softens the predictions of each model [19, 34]. In contrast to these methods which transfer supervision within the same modality, *cross-modal* distillation obtains supervision in one modality and transfers it to another. This approach was proposed for RGB and depth paired data, and for RGB and flow paired data by [31]. More recent work [3, 5, 6, 53] has explored this concept by exploiting the correspondence between synchronous audio and visual data in teacher-student style architectures [5, 6], or as a form of "self-supervision" [3] where networks for both modalities are learnt from scratch (an idea that was previously explored in the neuroscience community [9]). Some works have also examined cross-modal relationships between faces and voices in order to learn identity representations [39, 48, 49]. Differently from these works, our approach places an explicit reliance on the correspondence between the facial and vocal *emotions* emitted by a speaker during speech, discussed next.

Links between facial and vocal emotion. Our goal is to learn a representation that is aware of the emotional content in speech *prosody*, where prosody refers to the extra-linguistic variations in speech (e.g. changes in pitch, tempo, loudness, or intonation), by transferring such emotional knowledge from face images extracted synchronously. For this to be possible, the emotional content of speech must correlate with the facial expression of the speaker. Thus in contrast to multimodal emotion recognition systems which seek to make use of the complementary components of the signal between facial expression and speech [15], our goal is to perform cross-modal learning by exploiting the redundancy of the signal that is common to both modalities. Fortunately, given their joint relevance to communication, person perception, and behaviour more generally, interactions between speech prosody and facial cues have been intensively studied (Cvejic *et al.* [21]; Pell [56]; Swerts and Krahmer [61]). The broad consensus of these works is that during conversations, speech prosody is typically associated with other social cues like facial expressions or body movements, with facial expression being the most 'privileged' or informative stimulus [58].

Deep learning for speech emotion recognition. Deep networks for emotional speech recognition either operate on hand-crafted acoustic features known to have a significant effect on speech prosody, (e.g. MFCCs, pitch, energy, ZCR, ...), or operate on raw audio with little processing, e.g. only the application of Fourier transforms [20]. Those that use handcrafted features focus on global suprasegmental/prosodic features for emotion recognition, in which utterance level statistics are calculated. The main limitation of such global-level acoustic features is that they cannot describe the dynamic variation along an utterance [2]. Vocal emotional expression is shaped to some extent by differences in the temporal structure of language and emotional cues are not equally salient throughout

the speech signal [41, 58]. In particular, there is a well-documented propensity for speakers to elongate syllables located in word- or phrase-final positions [52, 55], and evidence that speakers vary their pitch in final positions to encode gradient acoustic cues that refer directly to their emotional state (Pell [55]). We therefore opt for the second strategy, using minimally processed audio represented by magnitude spectrograms directly as inputs to the network. Operating on these features can potentially improve performance "in the wild" where the encountered input can be unpredictable and diverse [40]. By using CNNs with max pooling on spectrograms, we encourage the network to determine the emotionally salient regions of an utterance.

Existing speech emotion datasets. Fully supervised deep learning techniques rely heavily on large-scale labelled datasets, which are tricky to obtain for emotional speech. Many methods rely on using actors [13, 14, 44, 47] (described below), and automated methods are few. Some video datasets are created using subtitle analysis [25]. In the facial expression domain, labels can be generated through reference events [1], however this is challenging to imitate for speech. A summary of popular existing datasets in given in Table 1. We highlight some common disadvantages of these datasets below, and contrast these with the VoxCeleb dataset that is used in this paper:

(1) Most speech emotion datasets consist of elicited or acted speech, typically created in a recording studio, where actors read from written text. However, as [27] points out, full-blown emotions very rarely appear in the real world and models trained on acted speech rarely generalise to natural speech. Furthermore there are physical emotional cues that are difficult to consciously mimic, and only occur in natural speech. In contrast, VoxCeleb consists of interview videos from YouTube, and so is more naturalistic.

(2) Studio recordings are also often extremely clean and do not suffer from 'real world' noise artefacts. In contrast, videos in the VoxCeleb dataset are degraded with real world noise, consisting of background chatter, laughter, overlapping speech and room acoustics. The videos also exhibit considerable variance in the quality of recording equipment and channel noise.

(3) For many existing datasets, cost and time prohibitions result in low speaker diversity, making it difficult to avoid speaker adaptation. Since our method does not require any emotion labels, we can train on VoxCeleb which is two orders of magnitude larger than existing public speech emotion datasets in the number of speakers.

Note that for any machine learning system that aims to perform emotion recognition using vision or speech, the ground truth emotional state of the speaker is typically unavailable. To train and assess the performance of models, we must ultimately rely on the judgement of human annotators as a reasonable proxy for the true emotional state of a speaker. Throughout this work we use the term "emotion recognition" to mean accurate prediction of this proxy.

3 CROSS MODAL TRANSFER

The objective of this work is to learn useful representations for emotion speech recognition, without access to labelled speech data. Our approach, inspired by the method of cross modal distillation [31], is to tackle this problem by exploiting readily available annotated data in the visual domain.

Under the formulation introduced in [31], a "student" model operating on one input modality learns to reproduce the features of a "teacher" model, which has been trained for a given task while operating on a different input modality (for which labels are available). The key idea is that by using a sufficiently large dataset of modality paired inputs, the teacher can transfer task supervision to the student without the need for labelled data in the student's modality. Importantly, it is assumed that the paired inputs possess the same attributes with respect to the task of interest.

In this work, we propose to use the correspondence between the emotion expressed by the facial expression of a speaker and the emotion of the speech utterance produced synchronously. Our approach relies on the assumption that there is some redundancy in the emotional content of the signal communicated through the concurrent expression and speech of a speaker. To apply our method, we therefore require a large number of *speaking face-tracks*, in which we have a known correspondence between the speech audio and the face depicted. Fortunately, this can be acquired, automatically and at scale using the recently developed SyncNet [18]. This method was used to generate the large-scale VoxCeleb dataset [50] for speaking face-tracks, which forms the basis of our study.

As discussed in Sec. 2, there are several ways to "distill" the knowledge of the teacher to the student. While [31] trained the student by regressing the intermediate representations at multiple layers in the teacher model, we found in practice that the approach introduced in [34] was most effective for our task. Specifically, we used a cross entropy loss between the outputs of the networks after passing both both sets of predictions through a softmax function with temperature T to produce a distribution of predictions:

$$p_i = \frac{\exp(x_i/T)}{\sum_j \exp(x_j/T)}, \tag{1}$$

where x_i denotes the logit associated with class i and p_i denotes the corresponding normalised prediction. A higher temperature softmax produces a "softer" distribution over predictions. We experimented with several values of T to facilitate training and found, similarly to [34], that a temperature of 2 was most effective. We therefore use this temperature value in all reported experiments.

3.1 The Teacher

This section describes how we obtain the teacher model which is responsible for classifying facial emotion in videos.

Frame-level Emotion Classifier. To construct a strong teacher network (which is tasked with performing emotion recognition from *face images*), training is performed in multiple stages. We base our teacher model on the recently introduced Squeeze-and-Excitation architecture [35] (the ResNet-50 variant). The network is first pretrained on the large-scale VGG-Face2 dataset [16] ($\approx$ 3.3 million faces) for the task of identity verification. The resulting model is then finetuned on the *FERplus* dataset [10] for emotion recognition. This dataset comprises the images from the original FER dataset ($\approx$ 35k images) [28] together with a more extensive set of annotations (10 human annotators per image). The emotions labelled in the dataset are: *neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, *fear* and *contempt*. Rather than training the teacher to predict a single correct emotion for each face, we instead require it to

Corpus	Speakers	Naturalness	Labelling method	Audio-visual
AIBO★ [11]	51	Natural	Manual	Audio only
EMODB [13]	10	Acted	Manual	Audio only
ENTERFACE [47]	43	Acted	Manual	✓
LDC [44]	7	Acted	Manual	Audio only
IEMOCAP [14]	10	Both†	Manual	✓
AFEW 6.0♠ [25]	unknown^{+}	Acted	Subtitle Analysis	✓
RML	8	Acted	Manual	✓
EmoVoxCeleb	**1,251**	**Natural**	**Expression Analysis**	✓

Table 1: Comparison to existing public domain speech emotion datasets. † contains both improvised and scripted speech. ★ contains only emotional speech of children. ♠ has not been commonly used for audio only classification, but is popular for audio-visual fusion methods. $^{+}$ identity labels are not provided.

Method	Accuracy (PrivateTest)
PLD [10]	85.1 ±0.5%
CEL [10]	84.6 ±0.4%
ResNet+VGG† [37]	87.4
SENet Teacher (Ours)	88.8 ±0.3%

Table 2: Comparison on the FERplus facial expression benchmark. † denotes performance of model ensemble. Where available, the mean and std. is reported over three repeats. The SENet Teacher model is described in Sec. 3.1.

match the *distribution* of annotator labels. Specifically, we train the network to match the distribution of annotator responses with a cross entropy loss:

$$\mathcal{L} = -\sum_n \sum_i p_i^{(n)} \log q_i^{(n)}, \tag{2}$$

where $p_i^{(n)}$ represents the probability of annotation n taking emotion label i, averaged over annotators, and $q_i^{(n)}$ denotes the corresponding network prediction.

During training, we follow the data augmentation scheme comprising affine distortions of the input images introduced in [63] to encourage robustness to variations in pose. To verify the utility of the resulting model, we evaluate on the FERPlus benchmark, following the test protocol defined in [10], and report the results in Table 2. To the best of our knowledge, our model represents the current state of the art on this benchmark.

From Frames to Face-tracks. Since a single speech segment typically spans many frames, we require labels at a face-track level in order to transfer knowledge from the face domain to the speech domain. To address the fact that our classifier has been trained on individual images, not with *face-tracks*, we take the simplest approach of considering a single face-track as a set of individual frames. A natural consequence of using still frames extracted from video, however, is that the emotion of the speaker is not captured with equal intensity in every frame. Even in the context of a highly emotional speech segment, many of the frames that correspond to transitions between utterances exhibit a less pronounced facial expression, and are therefore often labelled as 'neutral' (see Figure 2 for an example track). One approach that has been proposed to address this issue is to utilise a single frame or a subset of frames known as *peak frames*, which best represent the emotional content of the face-track [57, 64]. The goal of this approach is to select the frames for which the dominant emotional expression is at its apex. It is difficult to determine which frames are the key frames, however, while [57] select these frames manually, [64] add an extra training step which measures the 'distance' of the expressive face from the subspace of neutral facial expressions. This method also relies on the implicit assumption that all facial parts reach the peak point at the same time.

We adopt a simple approximation to peak frame selection by representing each track by the maximum response of each emotion across the frames in the track, an approach that we found to work well in practice. We note that prior work has also found simple average pooling strategies over frame-level predictions [8, 36] to be effective (we found average pooling to be slightly inferior, though not dramatically different in performance). To verify that max-pooling represents a reasonable temporal aggregation strategy, we applied the trained SENet Teacher network to the individual frames of the AFEW 6.0 dataset, which formed the basis of the 2016 Emotion Recognition in the Wild (EmotiW) competition [24]. Since our objective here is not to achieve the best performance by specialising for this particular dataset (but rather to validate the aggregation strategy for predicting tracks), we did not fine-tune the parameters of the teacher network for this task. Instead, we applied our network directly to the default face crops provided by the challenge organisers and aggregated the emotional responses over each video clip using max pooling. We then treat the predictions as 8-dimensional embeddings and use the AFEW training set to fit a single affine transformation (linear transformation plus bias), followed by a softmax, allowing us to account for the slightly different emotion categorisation (AFEW does not include a *contempt* label). By evaluating the resulting re-weighted predictions on the validation set we obtained an accuracy of 49.3% for the 7-way classification task, strongly outperforming the baseline of 38.81% released by the challenge organisers.

Figure 2: An example set of frames accompanying a single speech segment in the VoxCeleb dataset illustrating the *neutral transition-face* phenomenon exhibited by many face tracks: the facial expression of the speaker, as predicted by the static image-based face classifier often takes a 'neutral' label while transitioning between certain phonemes.

3.2 The Student

The student model, which is tasked with performing emotion recognition *from voices*, is based on the VGG-M architecture [17] (with the addition of batch normalization). This model has proven effective for speech classification tasks in prior work [50], and provides a good trade-off between computational cost and performance. The architectural details of the model are described in section 5.1.

3.3 Time-scale of transfer

The time-scale of transfer determines the length of the audio segments that are fed into the student network for transferring the logits from face to voice. Determining the optimal length of audio segment for which emotion is discernable is still an open question. Ideally, we would like to learn only features related to speech *prosody* and not the lexical content of speech, and hence we do not want to feed in audio segments that contain entire sentences to the student network. We also do not want segments that are too short, as this creates the risk of capturing largely neutral audio segments. Rigoulot, 2014 [58] studied the time course for recognising vocally expressed emotions on human participants, and found that while some emotions were more quickly recognised than others (fear as opposed to happiness or disgust), after four seconds of speech emotions were usually classified correctly. We therefore opt for a four second speech segment input. Where the entire utterance is shorter than four seconds, we use zero padding to obtain an input of the required length.

4 EMOVOXCELEB DATASET

We apply our teacher-student framework on the VoxCeleb [50] dataset, a collection of *speaking face-tracks*, or contiguous groupings of talking face detections from video. The videos in the VoxCeleb dataset are interview videos of 1,251 celebrities uploaded to YouTube, with over 100,000 utterances (speech segments). The speakers span a wide range of different ages, nationalities, professions and accents. The dataset is roughly gender balanced. The audio segments also contain speech in different languages. While the identities of the speakers are available, the dataset has *no emotion labels*, and the student model must therefore learn to reason about emotions entirely by transferring knowledge from the face network. The identity labels allow us to partition the dataset into three splits: `Train`, `Heard-Val` and `Unheard-Val`. The `Heard-Val` split contains held out speech segments from the same identities in the training set, while the `Unheard-Val` split contains identities

Figure 3: Examples of emotions in the EmoVoxCeleb dataset. We rely on the facial expression of the speaker to provide clues about the emotional content of their speech.

	Train	Heard-Val	Unheard-Val
# speaking face-tracks	118.5k	4.5k	30.5k

Table 3: The distribution of speaking face-tracks in the EmoVoxCeleb dataset. The `Heard-Val` set contains identities that are present in `Train`, while the identities in `Unheard-Val` are disjoint from `Train`.

that are disjoint from the other splits[2]. Validating on unheard identities allows us to ascertain whether the student model is exploiting identity as a bias to better match the predictions of the teacher model. The identity labels may also prove useful for researchers tackling other tasks, for example evaluating the effect of emotional speech on speaker verification, as done by [54]. The total size of each partition is given in Table 3.

By applying the teacher model to the frames of the VoxCeleb dataset as described in section 3.1, we automatically obtain emotion labels for the face-tracks and the speech segments. These labels take the form of a predicted distribution over eight emotional states that were used to train the teacher model: *neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, *fear* and *contempt*. These frame-level predictions can then be directly mapped to synchronous speech segments by aggregating the individual prediction distributions into a single eight-dimensional vector for each speech segment. For all experiments we perform this aggregation by max-pooling across frames. However, since the best way to perform this aggregation remains an open topic of research, we release the frame level predictions of the model as part of the dataset annotation. The result is a large-scale audio-visual dataset of human emotion, which we call the EmoVoxCeleb dataset. As a consequence of the automated labelling technique, it is reasonable to expect that the noise associated with the labelling will be higher than for a manually annotated

[2]The `Unheard-Val` split directly corresponds to the `Test (US-UH)` set defined in [48].

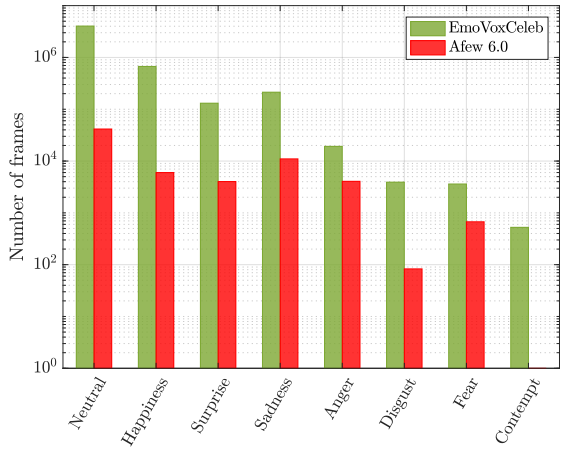

Figure 4: Distribution of frame-level emotions predicted by the SENet Teacher model for EmoVoxCeleb (note that the y-axis uses a log-scale). For comparison, the distribution of predictions are also shown for the Afew 6.0 dataset.

dataset. We validate our labelling approach by demonstrating quantitatively that the labels can be used to learn useful speech emotion recognition models (Sec. 5.2). Face-track visualisations can be seen in Figure 3, and audio examples are available online[3].

Distribution of emotions. As noted above, each frame of the dataset is annotated with a distribution of predictions. To gain an estimate of the distribution of emotional content in EmoVoxCeleb, we plot a histogram of the *dominant* emotion (the label with the strongest prediction score by the teacher model) for each extracted frame of the dataset, shown in Figure 4. While we see that the dataset is heavily skewed towards a small number of emotions (particularly neutral, as discussed in Sec. 3), we note that it still contains some diversity of emotion. For comparison, we also illustrate the distribution of emotional responses of the teacher model on 'Afew 6.0' [25], an emotion recognition benchmark. The Afew dataset was collected by selecting scenes in movies for which the subtitles contain highly emotive content. We see the distribution of labels is significantly more balanced but still exhibits a similar overall trend to EmoVoxCeleb. Since this dataset has been actively sampled to contain good diversity of emotion, we conclude that the coverage of emotions in EmoVoxCeleb may still prove useful, given that no such active sampling was performed. We note that Afew does not contain segments directly labelled with the *contempt* emotion, so we would therefore not expect there to be frames for which this is the predicted emotion. It is also worth noting that certain emotions are rare in our dataset. Disgust, fear and contempt are not commonly exhibited during natural speech, particularly in interviews and are therefore rare in the predicted distribution.

Data Format. As mentioned above, we provide logits (the pre-softmax predictions of the teacher network) at a frame level which can be used to directly produce labels at an utterance level (using max-pooling as aggregation). The frames are extracted from the face tracks at an interval of 0.24 seconds, resulting in a total of approximately 5 million annotated individual frames.

[3]http://www.robots.ox.ac.uk/~vgg/research/cross-modal-emotions

5 EXPERIMENTS

To investigate the central hypothesis of this paper, namely that it is possible to supervise a speech emotion recognition model with a model trained to detect emotion in faces, we proceed in two stages. First, as discussed in Sec. 4, we compute the predictions of the SENet Teacher model on the frames extracted from the VoxCeleb dataset. The process of distillation is then performed by randomly sampling segments of speech, each four seconds in duration, from the training partition of this dataset. While a fixed segment duration is not required by our method (the student architecture can process variable-length clips by dynamically modifying its pooling layer), it leads to substantial gains in efficiency by allowing us to batch clips together. We experimented with sampling speech segments in a manner that balanced the number of utterance level emotions seen by the student during training. However, in practice, we found that it did not have a significant effect on the quality of the learned student network and therefore, for simplicity, we train the student without biasing the segment sampling procedure.

For each segment, we require the student to match the response of the teacher network on the facial expressions of the speaker that occurred *during the speech segment*. In more detail, the responses of the teacher on each frame are aggregated through max-pooling to produce a single 8-dimensional vector per segment. As discussed in Section 3, both the teacher and student predictions are passed through a softmax layer before computing a cross entropy loss. Similarly to [34], we set the temperature of both the teacher and student softmax layers to 2 to better capture the confidences of the teacher's predictions. We also experimented with regressing the pre-softmax logits of the teacher directly with an Euclidean loss (as done in [7]), however, in practice this approach did not perform as well, so we use cross entropy for all experiments. As with the predictions made by the teacher, the distribution of predictions made by the student are dominated by the neutral class so the useful signal is primarily encoded through the relative soft weightings of each emotion that was learned during the distillation process. The student achieves a mean ROC AUC of 0.69 over the teacher-predicted emotions present in the unheard identities (these include all emotions except disgust, fear and contempt) and a mean ROC AUC of 0.71 on validation set of heard identities on the same emotions.

5.1 Implementation Details

The student network is based on the VGGVox network architecture described in [50], which has been shown to work well on spectrograms, albeit for the task of speaker verification. The model is based on the lightweight VGG-M architecture, however the fully connected *fc6* layer of dimension $9 \times n$ (support in both dimensions) is replaced by two layers – a fully connected layer of 9×1 (support in the frequency domain) and an average pool layer with support $1 \times n$, where n depends on the length of the input speech segment (for example for a 4 second segment, $n = 11$). This allows the network to achieve some temporal invariance, and at the same time keeps the output dimensions the same as those of the original fully

connected layer. The input to the teacher image is an RGB image,

Layer	Support	Filt dim.	# filts.	Stride	Data size
conv1	7×7	1	96	2×2	254×198
mpool1	3×3	-	-	2×2	126×99
conv2	5×5	96	256	2×2	62×49
mpool2	3×3	-	-	2×2	30×24
conv3	3×3	256	256	1×1	30×24
conv4	3×3	256	256	1×1	30×24
conv5	3×3	256	256	1×1	30×24
mpool5	5×3	-	-	3×2	9×11
fc6	9×1	256	4096	1×1	1×11
apool6	1×n	-	-	1×1	1×1
fc7	1×1	4096	1024	1×1	1×1
fc8	1×1	1024	1251	1×1	1×1

Table 4: The CNN architecture for the student network. The data size up until *fc6* is depicted for a 4-second input, but the network is able to accept inputs of variable lengths. Batch-norm layers are present after every conv layer.

cropped from the source frame to include only the face region (we use the face detections provided by the VoxCeleb dataset) resized to 224×224, followed by mean subtraction. The input to the student network is a short-term amplitude spectrogram, extracted from four seconds of raw audio using a Hamming window of width 25ms and step (hop) 10ms, giving spectrograms of size 512×400. At train-time, the four second segment of audio is chosen randomly from the entire speaking face-track, providing an effective form of data augmentation. Besides performing mean and variance normalisation on every frequency bin of the spectrogram, no other speech-specific processing is performed, e.g. silence removal, noise filtering, etc. (following the approach outlined in [50]). While randomly changing the speed of audio segments can be useful as a form of augmentation for speaker verification [50], we do no such augmentation here since changes in pitch may have a significant impact on the perceived emotional content of the speech.

Training Details. The network is trained for 50 epochs (one epoch corresponds to approximately one full pass over the training data where a speech segment has been sampled from each video) using SGD with momentum (set to 0.9) and weight decay (set to 0.0005). The learning rate is initially set to $1E-4$, and decays logarithmically to $1E-5$ over the full learning schedule. The student model is trained from scratch, using Gaussian-initialised weights. We monitor progress on the validation set of unheard identities, and select the final model to be the one that minimises our learning objective on this validation set.

5.2 Results on external datasets

To evaluate the quality of the audio features learned by the student model, we perform experiments on two benchmark speech emotion datasets.

RML: The RML emotion dataset is an acted dataset containing 720 audiovisual emotional expression samples with categorical labels: *anger, disgust, fear, happiness, sadness and surprise*. This database is language and cultural background independent. The video samples were collected from eight human subjects, speaking six different languages (English, Mandarin, Urdu, Punjabi, Persian, Italian). To further increase diversity, different accents of English and Chinese were also included.

eNTERFACE [47]: The eNTERFACE dataset is an acted dataset (in English) recorded in a studio. Forty-two subjects of fourteen nationalities were asked to listen to six successive short stories, each of which was designed to elicit a particular emotion. The emotions present are identical to those found in the RML dataset.

Both external datasets consist of acted speech, and are labelled by human annotators. Since the external datasets are obtained in a single recording studio, they are also relatively clean, in contrast to the noisy segments in EmoVoxCeleb. We choose the RML dataset for evaluation specifically to assess whether our embeddings can generalise to multilingual speech. Both datasets are class-balanced.

Method	RML		eNTERFACE	
	Modality	Acc.	Modality	Acc.
Random	A	16.7	A	16.7
Student	A	49.7 ± 5.4	A	34.3 ± 4.0
Teacher	V	72.6 ± 3.9	V	48.3 ± 4.9
Noroozi et al. [51]	A	65.3	A	47.1

Table 5: Comparison of method accuracy on RML and eNTERFACE using the evaluation protocol of [51]. Where available, the mean ± std. is reported.

We do not evaluate the predictions of the student directly, for two reasons: first, the set of emotions used to train the student differ from those of the evaluation test set, and second, while the predictions of the student carry useful signal, they skew towards neutral as a result of the training distribution. We therefore treat the predictions as 8-dimensional embeddings and adopt the strategy introduced in Sec. 3.1 of learning a map from the set of embeddings to the set of target emotions, allowing the classifier to re-weight each emotion prediction using the class confidences produced by the student. In more detail, for each dataset, we evaluate the quality of the student model embeddings by learning a single affine transformation (comprising a matrix multiply and a bias) followed by a softmax to map the 8 predicted student emotions to the target labels of each dataset. Although our model has been trained using segments of four seconds in length, its dynamic pooling layer allows it to process variable length segments. We therefore use the full speech segment for evaluation.

To assess the student model, we compare against the following baselines: the expected performance at chance level by a random classifier; and the performance of the teacher network, operating on the faces modality. We also compare with the recent work of [51], whose strongest speech classifier consisted of a random forest using a combination of 88 audio features inc. MFCCs, Zero Crossings Density (ZCD), filter-bank energies (FBE) and other pitch/intensity-related components. We report performance using 10-fold cross validation (to allow comparison with [51]) in Table 5. While it falls short of the performance of the teacher, we see that the student model performs significantly better than chance. These results indicate that, while challenging, transferring supervision from the facial domain to the speech domain is indeed possible. Moreover, we note that the conditions of the evaluation datasets differ significantly from those on which the student network was trained. We discuss this domain transfer problem for emotional speech in the following section.

5.3 Discussion

Evaluation on external corpora: Due to large variations in speech emotion corpora, speech emotion models work best if they are applied under circumstances that are similar to the ones they were

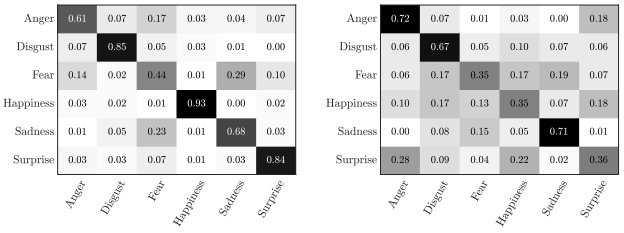

Figure 5: Normalised confusion matrices for the teacher model (left) and the student model (right) on the RML dataset (ground truth labels as rows, predictions as columns).

trained on [60]. For cross-corporal evaluation, most methods rely heavily on domain transfer learning or other adaptation methods [22, 23, 65]. These works generally agree that cross-corpus evaluation works to a certain degree only if corpora have similar contexts. We show in this work that the embeddings learnt on the EmoVoxCeleb dataset can generalise to different corpora, even with differences in nature of the dataset (natural versus acted) and labelling scheme. While the performance of our student model falls short of the teacher model that was used to supervise it, we believe this represents a useful step towards the goal of learning useful speech emotion embeddings that work on multiple corpora without requiring speech annotation.

Challenges associated with emotion distillation: One of the key challenges associated with the proposed method is to achieve a consistent, high quality supervisory signal by the teacher network during the distillation process. Despite reaching state-of-the-art performance on the FERplus benchmark, we observe that the teacher is far from perfect on both the RML and eNTERFACE benchmarks. In this work, we make two assumptions: the first is that distillation ensures that even when the teacher makes mistakes, the student can still benefit, provided that there is signal in the uncertainty of the predictions. The second is a broader assumption, namely that deep CNNs are highly effective at training on large, noisy datasets (this was recently explored in [45, 59], who showed that despite the presence of high label noise, very strong features can be learned on large datasets). To better understand how the knowledge of the teacher is propagated to the student, we provide confusion matrices for both models on the RML dataset in Figure 5. We observe that the student exhibits reasonable performance, but makes more mistakes than the teacher for every emotion except sadness and anger. There may be several reasons for this. First, EmoVoxCeleb used to perform the distillation may lack the distribution of emotions required for the student to fully capture the knowledge of the teacher. Second, it has been observed that certain emotions are easier to detect from speech than faces, and vice versa [15], suggesting that the degree to which there is a redundant emotional signal across modalities may differ across emotions.

Limitations of using interview data: Speech as a medium is intrinsically oriented towards another person, and the natural contexts in which to study it are interpersonal. Interviews capture these interpersonal interactions well, and the videos we use exhibit real world noise. However, while the interviewees are not asked to act a specific emotion, i.e. it is a 'natural' dataset, it is likely that celebrities do not act entirely naturally in interviews. Another drawback is the heavily unbalanced nature of the dataset where some emotions such as contempt and fear occur rarely. This is an unavoidable artefact of using real data. Several works have shown that the interpretation of certain emotions from facial expressions can be influenced to some extent by contextual clues such as body language [4, 33]. Due to the "talking-heads" nature of the data, this kind of signal is typically not present in interview data, but could be incorporated as clues into the teacher network.

Student Shortcuts: The high capacity of neural networks can sometimes allow them to solve tasks by taking "shortcuts" by exploiting biases in the dataset [26]. One potential for such a bias in EmoVoxCeleb is that interviewees may often exhibit consistent emotions which might allow the student to match the teacher's predictions by learning to recognise the identity, rather than the emotion of the speaker. As mentioned in Sec. 5, the performance of the student on the `heardVal` and `unheardVal` splits is similar (0.71 vs 0.69 mean ROC AUC on a common set of emotions), providing some confidence that the student is not making significant use of identity as a shortcut signal.

Extensions/Future Work: First, we note that our method can be applied as is to other mediums of unlabelled speech, such as films or TV shows. We hope to explore unlabelled videos with a greater range of emotional diversity, which may help to improve the quality of distillation and address some of the challenges discussed above. Second, since the act of speaking may also exert some influence on the facial expression of the speaker (for example, the utterance of an "o" sound could be mistaken for surprise) we would also like to explore the use of proximal *non-speech* facial expressions as a supervisory signal in future work. Proximal supervision could also address the problem noted in Section 3, that speaking expressions can tend towards neutral. Finally, facial expressions in video can be learnt using self-supervision [62], and this offers an alternative to the strong supervision used for the teacher in this paper.

6 CONCLUSIONS

We have demonstrated the value of using a large dataset of emotion unlabelled video for cross-modal transfer of emotions from faces to speech. The benefit is evident in the results – the speech emotion model learned in this manner achieves reasonable classification performance on standard benchmarks, with results far above random. We also achieve state of the art performance on facial emotion recognition on the FERPlus benchmark (supervised) and set benchmarks for cross-modal distillation methods for speech emotion recognition on two standard datasets, RML and eNTERFACE.

The great advantage of this approach is that video data is almost limitless, being freely available from YouTube and other sources. Future work can now consider scaling up to larger unlabelled datasets, where a fuller range of emotions should be available.

Acknowledgements. The authors would like to thank the anonymous reviewers, Almut Sophia Koepke and Judith Albanie for useful suggestions. We gratefully acknowledge the support of EPSRC CDT AIMS grant EP/L015897/1, and the Programme Grant Seebibyte EP/M013774/1.

REFERENCES

[1] S. Albanie and A. Vedaldi. Learning grimaces by watching tv. In *Proc. BMVC.*, 2016.

[2] Z. Aldeneh and E. M. Provost. Using regional saliency for speech emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2741–2745. IEEE, 2017.

[3] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617. IEEE, 2017.

[4] H. Aviezer, S. Bentin, R. R. Hassin, W. S. Meschino, J. Kennedy, S. Grewal, S. Esmail, S. Cohen, and M. Moscovitch. Not on the face alone: perception of contextualized face expressions in huntington's disease. *Brain*, 132(6):1633–1644, 2009.

[5] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[6] Y. Aytar, C. Vondrick, and A. Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[7] J. Ba and R. Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

[8] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 433–436. ACM, 2016.

[9] H. B. Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

[10] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

[11] A. Batliner, C. Hacker, S. Steidl, E. Nøth, S. D'Arcy, M. J. Russell, and M. Wong. You stupid tin box-children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In *LREC*, 2004.

[12] C. Bucilua, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541. ACM, 2006.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[14] 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*, 42(4):335, 2008.

[15] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[16] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018.

[17] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC.*, 2014.

[18] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[19] E. J. Crowley, G. Gray, and A. Storkey. Moonshine: Distilling with cheap convolutions. *arXiv preprint arXiv:1711.02613*, 2017.

[20] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller. An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 478–484. ACM, 2017.

[21] E. Cvejic, J. Kim, and C. Davis. Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication*, 52(6):555–564, 2010.

[22] J. Deng, Z. Zhang, F. Eyben, and B. Schuller. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072, 2014.

[23] J. Deng, Z. Zhang, and B. Schuller. Linked source and target domain subspace feature transfer learning–exemplified by speech emotion recognition. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 761–766. IEEE, 2014.

[24] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 427–432. ACM, 2016.

[25] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34–41, 2012.

[26] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[27] E. Douglas-Cowie, R. Cowie, and M. Schröder. A new emotion database: considerations, sources and scope. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[28] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[29] C. C. Goren, M. Sarty, and P. Y. Wu. Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, 56(4):544–549, 1975.

[30] T. Grossmann. The development of emotion perception in face and voice during infancy. *Restorative neurology and neuroscience*, 28(2):219–236, 2010.

[31] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2827–2836. IEEE, 2016.

[32] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 890–897. ACM, 2017.

[33] R. R. Hassin, H. Aviezer, and S. Bentin. Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review*, 5(1):60–65, 2013.

[34] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[35] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proc. CVPR*, 2018.

[36] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 553–560. ACM, 2017.

[37] C. Huang. Combining convolutional neural networks for emotion recognition. In *Undergraduate Research Technology Conference (URTC), 2017 IEEE MIT*, pages 1–4. IEEE, 2017.

[38] M. H. Johnson, S. Dziurawiec, H. Ellis, and J. Morton. Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40(1-2):1–19, 1991.

[39] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik. On learning associations of faces and voices. *arXiv preprint arXiv:1805.05553*, 2018.

[40] 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 2017 ACM on Multimedia Conference*, pages 1006–1013. ACM, 2017.

[41] Y. Kim and E. M. Provost. Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 92–99. ACM, 2016.

[42] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps. Transfer learning for improving speech emotion classification accuracy. *arXiv preprint arXiv:1801.06353*, 2018.

[43] J. Li, R. Zhao, J.-T. Huang, and Y. Gong. Learning small-size dnn with output-distribution-based criteria. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[44] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell. Ldc emotional prosody speech transcripts database. *University of Pennsylvania, Linguistic data consortium*, 2002.

[45] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. *arXiv preprint arXiv:1805.00932*, 2018.

[46] S. Mariooryad and C. Busso. Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 85–90. IEEE, 2013.

[47] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The enterfaceâĂŹ05 audio-visual emotion database. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 8–8. IEEE, 2006.

[48] A. Nagrani, S. Albanie, and A. Zisserman. Learnable PINs: Cross-modal embeddings for person identity. *Proc. ECCV*, 2018.

[49] A. Nagrani, S. Albanie, and A. Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proc. CVPR*, 2018.

[50] A. Nagrani, J. S. Chung, and A. Zisserman. VoxCeleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[51] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, 2017.

[52] D. K. Oller. The effect of position in utterance on speech segment duration in english. *The journal of the Acoustical Society of America*, 54(5):1235–1247, 1973.

[53] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, pages 801–816. Springer, 2016.

[54] S. Parthasarathy, C. Zhang, J. H. Hansen, and C. Busso. A study of speaker verification performance with expressive speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5540–5544. IEEE, 2017.

[55] M. D. Pell. Influence of emotion and focus location on prosody in matched statements and questions. *The Journal of the Acoustical Society of America*, 109(4):1668–1680, 2001.

[56] M. D. Pell. Prosody–face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior*, 29(4):193–215, 2005.

[57] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang. Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, 63:104–116, 2015.

[58] S. Rigoulot and M. D. Pell. Emotion in the voice influences the way we scan emotional faces. *Speech Communication*, 65:36–49, 2014.
[59] D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.
[60] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, 2010.
[61] M. Swerts and E. Krahmer. Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, 36(2):219–238, 2008.
[62] O. Wiles, A. S. Koepke, and A. Zisserman. Self-supervised learning of a facial attribute embedding from video. In *British Machine Vision Conference (BMVC)*, 2018.
[63] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM, 2015.
[64] S. Zhalehpour, Z. Akhtar, and C. E. Erdem. Multimodal emotion recognition based on peak frame selection from video. *Signal, Image and Video Processing*, 10(5):827–834, 2016.
[65] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller. Unsupervised learning in cross-corpus acoustic emotion recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 523–528. IEEE, 2011.
[66] S. Zhao, G. Ding, Y. Gao, and J. Han. Learning visual emotion distributions via multi-modal features fusion. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 369–377. ACM, 2017.